WHY BOTHER ME AGAIN? BENCHMARKING PERSONAL ASSISTANT MEMORY FOR INTERACTION PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Personal intelligence systems are taking on increasingly complex everyday tasks, yet repeated questions, unnecessary confirmations, and unwanted outreach can turn capable assistance into interaction burden. Reducing this burden requires assistants to remember how users want to be assisted, but existing memory benchmarks focus on factual recall and personalized answers and rarely test how assistants learn and act on interaction preferences. We introduce MemPA, an interactive environment for on-policy evaluation of memory in personal assistants. MemPA poses long-horizon tasks requiring assistants to decide when to clarify, confirm, or act autonomously. Four grounded simulated users hold 14 preferences across work and personal contexts. Their feedback spans 99 sessions over four months, with five event-driven changes per user testing memory updates. We evaluate preference selection, delivered behavior, and sensitivity to context and preference change. Evaluating four memory systems across four LLM backbones, we find that current memory brings limited gains in preference selection. Retrieval-based systems (i.e. Mem0, MemOS, simple RAG), do not outperform a memoryless assistant, while File-based memory improves accuracy but applies preferences beyond relevant contexts. Our analysis shows that current memory struggles to retrieve interaction preferences and adapt to preference changes. Even with the correct preference available, overriding the assistant's default interaction habits remains challenging. The environment and dataset are available here.

1 INTRODUCTION

AI assistants are moving from one-off task tools toward persistent personal assistants (PAs) that act on a user's behalf across messaging, files, scheduling, and long-running workflows, such as OpenClaw (OpenClaw Foundation, 2026). Task success and good interaction are separate outcomes: an agent can finish a task correctly while confirming too often, hiding its reasoning, or moving at an unwanted pace (Li et al., 2026a; Kirmayr et al., 2026). For a PA embedded in daily life, repeated questions, unnecessary confirmations, and unwanted outreach create a persistent interaction burden (Horvitz, 1999; McFarlane, 2002; Mehrotra et al., 2016; Zou et al., 2023b; Abbas et al., 2026). We call the preferences governing these behaviors *interaction preferences*: how the PA should talk, how much reasoning it should show, how much it should do autonomously, and how it should ask for information. Personalization research has mostly targeted content preferences, such as facts, topics, and choices a user favors, asking whether an assistant can recall or reflect them in its answers and decisions (Salemi et al., 2024; Wu et al., 2025a; Zhao et al., 2025; Jiang et al., 2025; Cheng et al., 2026; Chen et al., 2026). Interaction preferences shape every exchange, so getting them wrong imposes a cost even when the task succeeds.

PAs make interaction preferences harder to meet in two ways. First, interaction preferences depend on context. Users rely on the same assistant for work and personal matters (Chatterji et al., 2025), and the same behavior carries different risk in each: a user can want the PA to act on its own for work tasks and still require confirmation before sending a message to family (Parasuraman et al., 2000; Meurisch et al., 2020; Zhang et al., 2024; Pochampally et al., 2026). Second, interaction preferences change over time. An unwanted consequence of the PA's action, such as a message sent too early, can lead the user to revise how they want to be assisted, for example by requiring confirmation thereafter (Lee & Moray, 1992; Dietvorst et al., 2015; Baughan et al., 2023; Pochampally et al., 2026). Because

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.

a PA serves the same person for months, a preference applied in the wrong context or kept after it changes turns a single error into a burden that recurs across interactions. The PA therefore has to learn each preference from past interactions, apply it only in its context, and revise it when it changes. All three demands fall on the PA's memory.

Memory has to turn a long, noisy interaction history into preference evidence the PA can use at decision time. Interaction preferences make this hard. Users seldom state them outright; they reveal them through reactions to the PA, such as a correction or tolerated shortcut, scattered across unrelated sessions (Liu et al., 2025; Shi et al., 2026; Peng et al., 2026). A memory can store these reactions and still fail to bring the relevant one into the PA's input when needed (Wu et al., 2025a). The PA then falls back on its backbone's default interaction behavior. This default is shaped for task completion and an average user (Yao et al., 2024; Sorensen et al., 2024): models favor long responses and resolve ambiguity by guessing (Singhal et al., 2024; Zhang et al., 2025; Wu et al., 2025b). The resulting errors are systematic: the PA repeats the same default for users whose preferences differ from it.

Evaluating these demands requires a setting where interaction preferences surface through the user's reactions to the PA's behavior, accumulate over a persistent history shaped by the evaluated PA, differ across contexts, and change after events, while the PA acts on an executable world. Existing benchmarks each provide part of this setting (Table 1). AMemGym lets the evaluated assistant participate in a persistent user's history (Cheng et al., 2026), and VitaBench 2.0 adds executable tasks and preference changes (Chen et al., 2026); both target content and decision preferences. PrefIx and BenchPreS target interaction and communication preferences, but neither has a persistent user whose history the evaluated model shapes (Li et al., 2026a; Yoon et al., 2026).

We present MemPA, an executable environment for on-policy evaluation of whether a PA's memory can learn, separate, and update interaction preferences across context and time. MemPA follows four persona-grounded users through persistent, executable interactions with controlled preference changes and observable decision points, enabling evaluation of preference selection, behavioral realization, and user perception. Across four memory systems and four LLM backbones, we find that current memory remains insufficient for reliably adapting a PA's interaction behavior to the user's evolving preferences. Our contributions are as follows.

- We formulate interaction preferences as a state indexed by user, context, attribute, and time, and evaluate a PA's memory in three stages: acquiring a preference, selecting a setting, and realizing it in the delivered interaction.
- We build MemPA, an interactive environment where a PA goes through 99 on-policy sessions with each of four persona-grounded users, experiences controlled preference changes, and is evaluated on preference selection (PSA, CS, PET), delivered behavior (BRS), and user perception (PRS).
- We experiment on 4 memory systems with 4 LLM backbones, reveal that retrieval-based systems rarely surface applicable interaction preferences at decision time, while File memory improves selection but often applies preferences outside their context. We release the environment and data.

2 Related Work

Memory for Personalized Assistants Personalized assistants require memory systems to preserve user information and preferences from prior conversations (Wu et al., 2025a) and interactions (Ai et al., 2025), then tailor responses and decisions to the user (Li et al., 2025a). Specifically, memories should accurately identify preference-related information (Wu et al., 2025a), store and use it in assistance (Zhao et al., 2025), and update records when preferences change (Cheng et al., 2026; Chen et al., 2026; Liu et al., 2026). Benchmarks measure these capabilities through factual recall, personalized question answering, or recommendation (Wu et al., 2025a; Jiang et al., 2025; Zhao et al., 2025); Table 1 compares representative benchmarks along six dimensions. Recent benchmarks also examine whether personalized answers and task decisions reflect changes in user states and preferences (Cheng et al., 2026; Chen et al., 2026; Liu et al., 2026). PrefEval evaluates explicit and implicit content preferences through fixed multi-session histories and controlled conflicts where the scored target preference appears later (Zhao et al., 2025). A separate line of work evaluates adherence to interaction preferences (Li et al., 2026a) and contextual appropriateness of communication styles (Yoon et al., 2026). While useful for assessing personalized answers, task decisions, and preference adaptation, these benchmarks offer limited insight into how memory supports behavioral adaptation as interaction preferences vary across contexts and change over time. Here, we define memory

requirements for interaction preferences by their contextual breadth across settings and temporal depth over long-term interactions.

Table 1: Comparison of MemPA with representative personalization benchmarks for agent memory.

Work	Personalization target	Executable task env. (E)	Preference-conditioned reactions (R)	Persistent-user on-policy histories (H)	Contextual isolation (I)	Controlled change (C)
PrefIx	Interaction preference	✓	✓	—	—	—
AMemGym	Content / user state	—	◐	✓	—	◐
VitaBench 2.0	Content / decision preference	✓	—	◐	◐	✓
BenchPreS	Communication / style	—	—	—	✓	—
MemPA	**Interaction preference**	✓	✓	✓	✓	✓

Legend: ✓ supported; ◐ partially supported; — not covered by the native protocol. MemPA histories combine live accumulation sessions with fixed change events.

Interactive Executable Environments and User Simulation Executable environments are widely used to evaluate agents on realistic, multi-step tasks (Ding et al., 2026; Yao et al., 2024; Trivedi et al., 2024). Given intermediate results and changing state, a PA must decide how to involve the user, such as clarifying an underspecified goal, seeking confirmation before an action, or finding an alternative when the intended action is infeasible. Such environments are therefore a natural testbed for interaction intelligence. However, this also requires a user simulator that gives consistent reactions beyond task-level feedback. Existing simulated users are often loosely defined or built from attribute-assembled profiles, which may support population-level studies but not individual interaction preferences (Ai et al., 2025; Duan et al., 2026; Zhou et al., 2024; Li et al., 2026b). Grounding a character in detailed source material and a concrete world makes role-play more consistent and better grounded (Zong et al., 2026). MemPA therefore grounds users in detailed personas and personal worlds, where reactions reflect both the PA's behavior and the user's circumstances. Moreover, memory-driven personalization requires on-policy interaction, where preferences are learned from consequences caused by the evaluated assistant rather than from preconstructed histories (Cheng et al., 2026). MemPA combines these elements to evaluate whether a PA can learn and apply user-specific interaction preferences over long-horizon interactions.

3 MemPA: Long-Horizon Memory Benchmark for PA

Figure 1 gives an overview of MemPA. We formalize interaction preferences and how the user, world, PA, and memory interact (Section 3.1), build the users, worlds, and timeline (Section 3.2), connect the simulated user and the PA (Section 3.3), and score preference selection and realization with probes (Section 3.4).

3.1 Problem Formulation: Dynamic Interaction Preferences

We model a user's interaction preferences as a state indexed by context, attribute, and time. For user u, context $c \in \mathcal{C}$, and interaction attribute a, the target preference $p^{(t)}_{u,c,a} \in \mathcal{P}_a$ is one of the discrete settings of a at turn t of the user's full timeline. With $\mathcal{C} = \{\text{work}, \text{personal}\}$ and 14 attributes, each user has 28 context–attribute cells, whose settings at turn t we write as $p^{(t)}_u$. Each user lives in a persistent world that evolves through interaction with the PA:

$$x_t \sim U_u\big(\cdot \mid W_t,\, y_{t-1},\, S_t;\, p^{(t)}_u\big), \qquad W_{t+1} = \mathcal{T}(W_t, y_t), \qquad M_{t+1} = \mathcal{A}(M_t, x_t, y_t). \quad (1)$$

The user writes each message x_t from the world state W_t, the PA's previous interaction y_{t-1}, and its own notes S_t on earlier sessions, guided by $p^{(t)}_u$. The tool actions in the PA's delivered interaction y_t change the world through $\mathcal{T}$, so later tasks and reactions build on what this PA did. The evaluated memory system writes each exchange through its own mechanism $\mathcal{A}$, so its state M_t is built from the history $H_{<t} = (x_1, y_1, \ldots, x_{t-1}, y_{t-1})$. The settings stay fixed except at five controlled events, each of which changes one cell (Section 3.2).

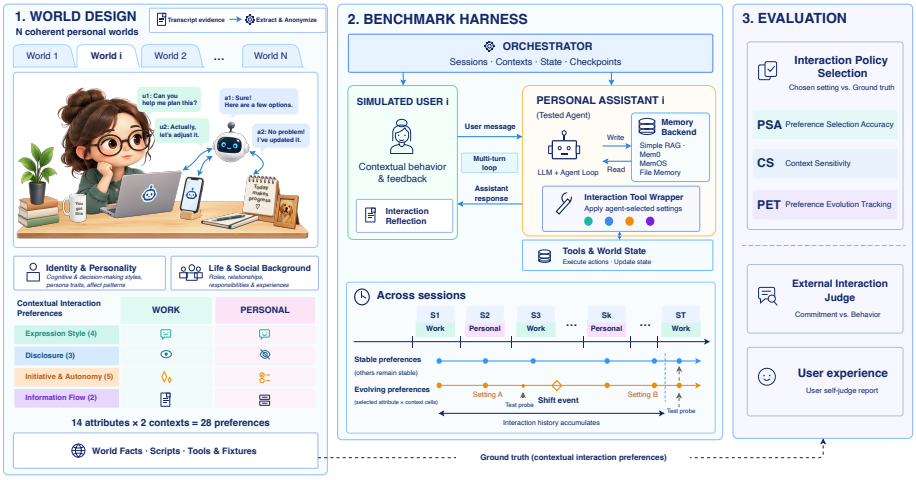


Figure 1: Overview of the benchmark: persona-grounded world design, a multi-session interaction harness connecting the simulated user and personal assistant, and evaluation of interaction preferences.

At each turn, the PA turns what its memory holds into the interaction it delivers:

$$E_t = \mathcal{R}(M_t, x_t), \qquad \big(\hat{p}^{(t)}_{u,c,a},\, \tilde{y}_t\big) = \pi(x_t, E_t), \qquad y_t = \Omega\big(\tilde{y}_t, \hat{p}^{(t)}_{u,c,a}\big). \tag{2}$$

The access mechanism $\mathcal{R}$ of the evaluated system returns evidence E_t. From x_t, the earlier turns of the current session, and E_t, the policy π selects a setting $\hat{p}^{(t)}_{u,c,a}$ and prepares an interaction $\tilde{y}_t$ with its reply and proposed tool actions, which the IX Wrapper Ω carries out according to the selection (Section 3.3). The selection is tested at a decision point, a turn at which the task requires a choice about attribute a. There, every setting in $\mathcal{P}_a$ is feasible and leads to an observably different interaction, and the user does not name the setting, so no setting is correct for a PA that does not know the user: the PA can infer the target only from $H_{<t}$ through M_t (Section 3.2).

Equations (1) and (2) separate three stages at which a PA can fail. In *acquisition*, $\mathcal{A}$ must turn the user's reactions into preference knowledge in M_t; this happens inside each memory system in its own format, so we examine it through diagnostic cases. In *selection*, $\mathcal{R}$ and π must recover the current setting, $\hat{p}^{(t)}_{u,c,a} = p^{(t)}_{u,c,a}$, which PSA scores over all cells, CS over the attributes whose work and personal settings differ, and PET over the changed cells before and after their change. In *realization*, y_t must follow that setting, which BRS and PRS score (Section 3.4).

3.2 Users, Worlds, and Timeline

Interaction Attributes. We describe how a user wants to be assisted with 14 interaction attributes, each with a set of discrete settings $\mathcal{P}_a$. The attributes are adapted from PrefIx (Li et al., 2026a), which defines interaction preferences for single-task tool-use agents. A PA instead serves the same person for months and acts on their behalf across work and personal life, so we rewrite each attribute for this setting under PA's scenarios, such as tone, emotional engagement, and proactive outreach. Each attribute ties a user preference to an observable and actionable aspect of PA behavior. The attributes fall into four dimensions: Expression Style concerns how the PA talks, Disclosure how much of its reasoning and process it shows, Initiative & Autonomy how much it does on its own, and Information Flow how it asks for information and handles multiple topics. Appendix A gives the definitions and settings of all attributes, together with the cognitive and personality traits that psychology and HCI research link to each.

Persona-Grounded Users. A person's interaction preferences with a PA largely carry over from how they want others to help them (Nass & Brave, 2005), and these habits follow from the whole person (Cacioppo & Petty, 1982; Burger & Cooper, 1979; Webster & Kruglanski, 1994). We therefore ground each user in a behaviorally distinctive and richly documented television persona, drawing four users from *The Big Bang Theory* and the U.S. *The Office*, and derive each user's setting for every context-attribute cell from the full transcripts. We check every LLM extracted interaction preference against its cited scenes: the high confidence ones are confirmed, and the authors set cells that have no

usable transcript evidence (Table 5). Of the 56 user-attribute pairs, 18 take different settings at work and in personal life. Appendix B details the pipeline.

Anonymization. A PA could guess a well-known character's preferences from prior knowledge instead of from memory, so we anonymize each user by replacing their names, workplaces, and places while keeping their preference settings, personality, and language style. The anonymized users still react in line with their extracted preferences, and the shortcut disappears: a backbone that predicts a user's 28 settings reaches 53.2% accuracy from the source identity but 46.4% from the anonymized profile, close to the 45.2% of chance (Appendix B.4).

World and Tools. Each user u lives in a world with a storyline that covers four months of their work and personal life. The PA reads and changes the world state W_t through 83 executable tools across 26 domains (Table 7), and these tool calls implement the transition $\mathcal{T}$ (Appendix C).

Sessions and Decision Points. Each session targets one preference cell (c, a). Its task is seeded from one of the scenario types of context c and placed on a date in the storyline. A live dialogue can drift from the planned path, so the task itself makes the decision point come up. The user has an explicit goal, but part of the information needed to reach it is missing from both W_t and the opening request. The task sets the conditions for its completion so that they pass through the choice about a, and the user supplies them at a designated step, so the PA can finish the task only by making this choice (Appendix C.3). Rule-based programs test the fixtures and the tool logic, and human-verified LLM-based semantic checks confirm for each session that every setting in $\mathcal{P}_a$ is viable: the PA can carry it out with the records and tools in the world, it leads to an observably different interaction, and the user has a natural way to respond to it (Appendix C.5).

Timeline and Preference Change. Each user's timeline consists of 99 sessions over the four months: 94 accumulation sessions, where the PA learns preferences through live interaction, and five controlled preference-change events. Each of the 23 stable cells appears in three accumulation sessions, and each of the five changing cells appears in three sessions before its change and two after. In each event e_k, a PA action has a consequence that leads the user to revise the setting from $p^{(t_k^-)}_{u,c_k,a_k}$ to $p^{(t_k^+)}_{u,c_k,a_k}$; its transcript is written in advance and replayed through the same interface as a live session (Section 3.3), so every memory condition experiences the same change. Probes run at checkpoints along this timeline (Section 3.4).

3.3 Interaction Harness

User Simulator. A separate nanobot agent plays user U_u from the anonymized profile, current settings $p_u^{(t)}$, and the session script (Section 3.2). It sees the delivered interaction y_{t-1} and results of executed actions, but not the PA's selection or anything the wrapper holds back. It keeps its own memory S_t across sessions, including lasting impressions such as "it tends to over-explain", so its reactions accumulate like a person's and follow its personality: a user who corrected the same mistake twice may respond more urgently the third time, while a patient user may react calmly. It expresses preferences only through reactions to the PA's behavior; prompt rules, rule-based checks, and an LLM judge block any message that names a setting or leaks information (Section 4). Two annotators confirmed on a sample of accumulation sessions that these reactions convey the preference (99.3%), and a post-hoc audit at formal run found no leakage (Appendices F.9 and D.4).

PA and Memory Interface. The PA is built on nanobot (HKUDS, 2026), an open-source personal agent framework, and the evaluated LLM backbone serves as its policy π. The PA acts on the world through the tools of Section 3.2. Any memory system connects to the PA through a common adapter. After each turn, the adapter passes the turn as the PA saw it to the system's write mechanism $\mathcal{A}$: the user message x_t, the tool calls of y_t with their results, the PA's call to the interaction tool, and the reply text of y_t. It places the evidence E_t returned by the access mechanism $\mathcal{R}$ in the PA's prompt. The target setting and any content held back by the wrapper never reach the memory. Every memory system builds its own history on the same users, sessions, tools, and wrapper. Section 4 describes the four memory systems we evaluate.

IX Wrapper. Current LLM agent runtimes do not support many interaction preferences as executable controls, such as holding an action until user approval. MemPA therefore has the PA state its choice explicitly: at a decision point for attribute a, the PA must call its interaction tool (e.g., `IX_autonomy_level`) with one setting $\hat{p} \in \mathcal{P}_a$, and the IX Wrapper Ω carries it out by passing the PA's wording through or by controlling which part of $\tilde{y}_t$ is shown, when it is delivered, or whether a

proposed action runs. The wrapper never sees the target setting $p^{(t)}_{u,c,a}$, so it cannot correct a wrong choice. The attribute-to-mechanism mapping and a worked example are given in Appendix E.

3.4 Evaluation

Probes. MemPA scores selection and realization (Section 3.1) with probes. A probe is a short session with one decision point for one preference cell. Each user has 33 probes: 28 final probes cover every cell after the last session, and five pre-event probes test changing cells just before their events. Each probe starts from the checkpoint frozen at that point and runs on an isolated branch. The branch carries world state and memory M_t but no raw conversation from earlier sessions, so the PA can access its history only through memory. Any probe-induced change to the world or memory is discarded, so probes cannot affect one another. Since LLM agents vary across runs, each probe is repeated three times, and we report the mean for each backbone and memory condition. Failure handling and pooling of CS and PET are given in Appendix F.3.

Preference Selection. Selection is directly observable, since the PA states $\hat{p}^{(t)}_{u,c,a}$ through the interaction tool. For each probe, we score it by exact match, $z^{(t)}_{u,c,a} = \mathbf{1}\big[\hat{p}^{(t)}_{u,c,a} = p^{(t)}_{u,c,a}\big]$, and report three metrics. **Preference Selection Accuracy (PSA)** is the fraction of final probes with a correct selection. **Context Sensitivity (CS)** considers the user–attribute pairs whose settings differ between work and personal contexts, and counts a pair only when both selections are correct. For example, if the target is `Terse` at work and `Detailed` in personal contexts, the PA must select both. **Preference Evolution Tracking (PET)** considers the changing cells, and counts a cell only when the selections before and after its change are both correct:

$$\mathrm{PSA} = \tfrac{1}{|P|}\textstyle\sum_P z^{(\mathrm{final})}_{u,c,a}, \;\; \mathrm{CS} = \tfrac{1}{|D|}\textstyle\sum_D z^{(\mathrm{final})}_{u,\mathrm{work},a}\, z^{(\mathrm{final})}_{u,\mathrm{personal},a}, \;\; \mathrm{PET} = \tfrac{1}{|E|}\textstyle\sum_E z^{(\mathrm{pre})}_{u,c,a}\, z^{(\mathrm{final})}_{u,c,a}.$$

Here, P contains all final-probe cells, D contains the user–attribute pairs whose work and personal settings differ, and E contains the changing cells.

Behavioral Realization. A correct selection does not guarantee a well-followed delivered interaction, since the PA may still write or act against the setting it chose. We therefore score how well y_t follows the target setting $p^{(t)}_{u,c,a}$ with an LLM judge. The judge reads the PA's public reply and the actions that actually ran, together with the definition of the target setting, and gives a **Behavioral Realization Score (BRS)** from 1 to 5. The judge does not see the PA's call to the interaction tool or its selection $\hat{p}$. The judge prompt and criteria are given in Appendix G.4.

Perceived Realization. The user also rates the delivered interaction. After each final probe, the same user simulator, without accumulated memory, re-reads the probe from its side and rates, from 1 to 5, how well the PA met its preference on the probed attribute. We call this the **Perceived Realization Score (PRS)**. We use PRS to compare memory conditions run on the same users (Finding 4) and BRS to compare settings held by different users (Finding 7). BRS and PRS agree closely: across 7,996 final probes with both scores, Spearman $\rho = 0.83$ (Appendix F.9).

Reference Conditions. Two reference conditions bound the memory conditions and show at which stage the PA fails. *No Memory* skips accumulation and gives the PA no memory, so its selections and delivered interactions reflect the backbone's default tendencies. *Oracle* replaces the evidence E_t at each probe with the target setting $p^{(t)}_{u,c,a}$ and its definition, which removes acquisition from the pipeline. A remaining selection error then means the backbone does not follow a preference it has been told, and a low BRS after a correct selection means it cannot carry out the setting. The BRS under *Oracle* therefore serves as the realization reference for each backbone.

4 Experimental Setup

We run every combination of four PA backbones and four memory systems on all four users, and add two reference conditions for each backbone (Section 3.4).

PA Backbones. We evaluate GPT-5.6-Sol, Qwen3.8-27B, HY4-Preview, and DeepSeek-V4-Flash through the OpenRouter chat-completions API with tool calling. All four use temperature 0.7 and at most 8,192 output tokens per call. We do not set reasoning effort, so each model reasons at its default level. The PA can make up to 40 model calls per user turn. With the simulator and judge in the loop, one accumulation session takes about 10–20 minutes, and one probe under 5 minutes.

Memory Systems. We evaluate four systems. File memory, nanobot's built-in memory, uses an LLM to consolidate earlier conversation into `MEMORY.md` and gives the full file to the PA. Simple RAG stores raw messages and retrieves them by vector similarity. Mem0 v2.0.2 (Chhikara et al., 2025) extracts memories with an LLM and searches them, while MemOS v2.0.15 (Li et al., 2025b) stores and searches memories through its MemCube. All backbones share the same memory-side models (Table 11). In File memory, DeepSeek-V4-Flash consolidates the oldest turns into `MEMORY.md` once the PA prompt reaches about 20k tokens, until it drops below 10k; the PA reads the full file in its system prompt. Mem0 extracts memories with DeepSeek-V4-Flash. Simple RAG uses the same Mem0 store with extraction disabled (`infer=False`), storing raw messages. MemOS uses GPT-5.5 as its memory LLM and keeps only textual memory. The three retrieval-based conditions embed text with text-embedding-3-small, query memory with each user message as in Mem0's reference integration, and place at most 4,000 tokens from the top 10 results in the PA's system prompt. They trim raw conversation at the same 20k-token limit, dropping turns already stored in memory.

Simulator and Judges. Gemini-3.8-Flash plays the user, with temperature 0 and at most 4,096 output tokens per turn. Rule-based checks and a Gemma-4-31B judge review each simulator message before the PA receives it. They keep the conversation on the session's task until the decision point for the tested attribute arrives, and they stop the simulator from revealing, before that point, the information the decision point will ask for. Gemini-3.8-Flash gives both realization scores: as the judge, it gives BRS with one fixed prompt and rubric (Appendix G.4), and as the user simulator without its accumulated memory, it gives PRS after each final probe. Claude Sonnet 5, which serves neither as PA nor as simulator, re-judges a stratified sample of 418 probes as a second judge (Appendix F.9). Claude Opus 5.5 reads simulator sentences for the post-hoc leakage audit (Appendix D.4).

Reference Conditions. We run *No Memory* and *Oracle* with each backbone on the probes only, without accumulation. Each probe starts from the checkpoint of the same user's DeepSeek-V4-Flash File memory run, which provides the world state at the probe's point on the timeline, with the PA's memory cleared and no earlier conversation. Under *Oracle*, the target setting and its definition take the place of the retrieved memory in the PA's prompt.

5 Results and Analysis

Finding 1: Only File Memory Helps PSA. Current memory systems give the PA only a small advantage in selecting the user's interaction preferences. Table 2 reports PSA for every user, backbone, and memory condition. Averaged with equal weight over the four backbones, No Memory, where the PA selects by its defaults, reaches 49.6%, compared with 45.2% for random selection (Appendix A.1). Only File memory adds a clear gain. It raises PSA to 59.2%, a mean gain of 9.6 points over No Memory (95% cluster-bootstrap interval [+3.9, +15.6]). In PERMA, agents with Mem0, MemOS, or retrieval over past dialogue answer most questions about the user's content preferences correctly (Liu et al., 2026). On interaction preferences, Mem0, MemOS, and Simple RAG stay at the level of No Memory, with PSA of 50.0%, 50.0%, and 49.9%. The intervals of all conditions and their comparisons with No Memory on each user–backbone pair are given in Table 13 and Appendix F.4. The PSA ranking of the backbones does not follow their general capability (Table 19).

Table 2: Preference Selection Accuracy (PSA, %) across users, backbones, and memory conditions. Probes that end in a model failure score 0 (Appendix F.3).

Memory	User A				User B				User C				User D			
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash
File memory	60.7	75.0	52.4	64.3	60.7	72.6	69.0	57.1	48.8	42.9	51.2	47.6	61.9	59.5	63.1	60.7
Simple RAG	38.1	46.4	44.0	44.0	58.3	52.4	45.2	52.4	36.9	48.8	48.8	47.6	50.0	66.7	53.6	64.3
Mem0	40.5	50.0	40.5	38.1	56.0	53.6	53.6	57.1	38.1	47.6	41.7	41.7	59.5	66.7	59.5	56.0
MemOS	40.5	48.8	44.0	40.5	57.1	46.4	52.4	54.8	34.5	45.2	41.7	42.9	54.8	72.6	60.7	63.1
No Memory	41.7	41.7	47.6	42.9	54.8	57.1	54.8	54.8	39.3	44.0	42.9	45.2	53.6	67.9	47.6	58.3
Oracle	100.0	98.8	100.0	100.0	97.6	100.0	100.0	96.4	98.8	100.0	98.8	96.4	100.0	100.0	97.6	100.0

Each cell reports mean PSA (%) over three probe runs.

Finding 2: Content Yes, Interaction No. To see what each memory brings into the PA's input, we sampled 112 final probes per memory condition, seven for each user and backbone, and examined the input that the PA received before its selection on the target attribute. Two Claude Opus 5.5 readers

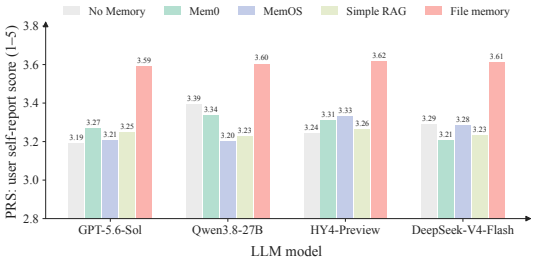


Figure 2: PRS across backbones and memory conditions, averaged over the 3 probe repeats and then over the 4 users. PRS is the user simulator's 1–5 rating of how well the PA met its preference on the target attribute, and Oracle scores are given in Table 16.

judged whether each input holds an applicable interaction preference, blind to the memory condition and backbone. We found that the content retrieved by the three retrieval-based conditions rarely contains such a preference (Figure 7 in the appendix). It appears in only 7 of the 334 judged probes, compared with 60 of 111 (54%) under File memory. Under these three conditions, the PA sees only content-related events in 55–73% of the judged probes.

Finding 3: Stored yet Not Retrieved. For Simple RAG, which stores the raw messages by design, a preference can only go missing from the input at retrieval. To also check the other two conditions, we inspected the stores of all three retrieval-based conditions for two final probes of User B, one on Proactive Outreach at work and one on Topic Management in the personal context. In these cases, the preference had been written to each store, yet none of the three conditions retrieves it before the PA's selection (Figure 8 in the appendix). On Proactive Outreach, File memory holds the user's objection to a reminder that the PA had sent unprompted, and the PA selects the target setting Low. Simple RAG returns that earlier reminder but leaves the objection in the store, and Mem0 and MemOS return content-related events. With all three, the PA wrongly selects High.

Finding 4: File Memory Also Delivers Better on PRS and BRS. File memory reaches a PRS of 3.59–3.62 across the four backbones, above the 3.20–3.34 of the three retrieval-based conditions and the 3.19–3.39 of No Memory (Figure 2). BRS, the LLM judge's score, shows the same pattern (Figure 9 in the appendix). File memory scores 3.46–3.57 and is the highest condition for each backbone, while the three retrieval-based conditions stay at 3.03–3.26, close to the 3.01–3.29 of No Memory. BRS shows no clear relation to the backbone, and within one memory condition the four backbones differ by less than 0.3. The agreement between BRS and PRS on each probe and a cross-check with a second judge are given in Appendix F.9. Per-user scores are reported in Tables 15 and 16, and per-dimension scores in Tables 17, 18.

Finding 5: Preferences Spill across Contexts in CS and PET. CS and PET stay low under every memory condition (Table 3). Each table entry pools the CS pairs or PET cells of User A-D within a repeat and averages the three repeats. We compare the memory conditions with No Memory (CS: 21.8% and PET: 11.2%) averaged over the four backbones. Mem0, MemOS, and RAG stay at the level of No Memory, with CS at 19.9–22.2% and PET at 10.8–15.0%. Interestingly, treating No Memory as the backbone's default interaction policy, we can see that most wrong selections in these pairs and cells match this default, and so do all eight wrong selections under Oracle. The backbones therefore carry a strong interaction prior. (Appendix F.7).

File memory follows a change but not a context. Its PET is 25.4%, more than twice the 11.2% of No Memory, and it is higher on every backbone. However, the CS does not improve, at 14.4% against 21.8%. (Appendix F.7). We see 26.9% of CS pairs are carry-overs. It keeps the preference but not the context it belongs to. In a personal session, User B asks the PA to work on one task at a time. File memory records this rule under a general protocol heading, and it records a later correction from a work session only as an exception for closely related sub-items. In a work probe on Topic Management whose target is Organize, the PA with File memory selects One-at-a-time in all three repeats on three of the four backbones.

Finding 6: Saying Is Easier than Doing across IX Dimensions. Across the four dimensions, Information Flow is the hardest under every condition (Figure 3). Its PSA stays at 34–44%, the lowest of the four, while Expression Style (57–69%) and Disclosure (55–58%) are relatively easier. A breakdown of the PSA of each IX dimension is given in Table 14. The difference may lie in what

these preferences really change. Every preference in Expression Style and Disclosure changes how the PA's reply reads, while most preferences in Information Flow and Initiative & Autonomy change what the PA does in the task, such as whether it asks the user before it acts or takes a step the user did not request. Preferences on the PA's actions are harder to select than preferences on its reply under every memory condition, and under No Memory they fall below random selection (Appendix F.5).

Table 3: CS and PET across backbones and memory conditions.

Memory	CS					PET				
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	Avg	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	Avg
File memory	14.8	13.0	22.2	7.4	14.4	21.7	30.0	38.3	11.7	25.4
Simple RAG	14.8	29.6	20.4	18.5	20.8	20.0	15.0	16.7	8.3	15.0
Mem0	25.9	29.6	22.2	11.1	22.2	18.3	16.7	15.0	6.7	14.2
MemOS	20.4	24.1	22.2	13.0	19.9	13.3	15.0	8.3	6.7	10.8
No Memory	27.8	27.8	22.2	9.3	21.8	11.7	15.0	13.3	5.0	11.2
Oracle	100.0	100.0	100.0	98.1	99.5	98.3	100.0	98.3	96.7	98.3

Finding 7: Realization Is Not Easy even under Oracle. Some settings still fall short even after the PA is told the target setting. We compare settings with BRS, which rates every reply with the same rubric for all users (Appendix G.4). Under Oracle, the PA receives the target setting, and we take as the pass rate of a target setting the share of its correctly selected final probes with a BRS of 4 or higher. Five settings have the lowest pass rates: Hide for Reasoning Visibility, Terse for Verbosity, Iterative and Structured for Information Elicitation, and Low for Solution Breadth. Together, their correctly selected probes pass in only 60% of cases, against 95% for the other 25 settings. The judge's rationales for these shortfalls show two tendencies of current models. On Hide, Terse, and Low, the **PA cannot help saying more than the setting allows**, adding reasons for its answer or options it did not choose (Singhal et al., 2024). On Iterative and Structured, **the PA mostly asks too little and acts on assumptions about missing facts** (Zhang et al., 2025; Wu et al., 2025b). These tendencies reflect the PA's default way of replying. Compared with No Memory, Oracle raises correct selections of the five settings from 33% to 96%, but the share of replies with a BRS of 4 or higher rises only from 24% to 59%. Even with the IX Wrapper, the PA still writes too much. After it selects Hide, it can still add a "Why that day:" section with three reasons, as Show would (Figure 10). More examples of correct selections whose delivered behavior falls short are given in Appendix F.8. Overcoming the PA's interaction habits is therefore harder than it may seem, suggesting that aligning models with users' interaction preferences requires more attention to these behaviors.

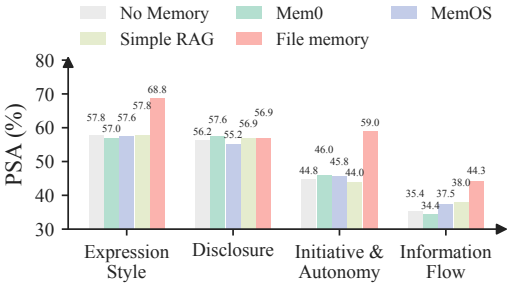


Figure 3: PSA by IX preference dimension and memory condition.

6 Discussion

In this paper, we formalized interaction preferences and built MemPA, an interactive environment for on-policy evaluation of whether personal assistants learn how a user wants to be assisted and apply it across contexts and as preferences evolve. Evaluating four memory systems on four diverse and richly characterized users, we show that current memory brings little gain in preference selection. The retrieval-based systems do not surpass the memoryless level. File memory improves preference selection and better captures preference changes, yet still fails to constrain preferences to the contexts where they apply. Our diagnostic analysis further shows that the retrieval-based systems can store an interaction preference and still fail to retrieve it at the decision. The two reference conditions, No Memory and Oracle, further reveal that overriding the model's default way of interacting is hard, even for a memory that holds the right preference. Our findings raise two questions for the ongoing discussion of agent personalization with memory: (1) how memory systems can support alignment with users' subtler interaction-level preferences, and (2) what memory systems can do to keep personalization within the right range as contexts differ and preferences shift.

ETHICS STATEMENT

Purpose. MemPA is intended to support assistants that adapt to how each user wants to be assisted, so that people spend less effort repeating instructions, answering unnecessary questions, and correcting unwanted actions. The benchmark scores agreement with each user's own current preferences, including after those preferences change, rather than engagement or task throughput.

Human subjects and privacy. No human participants took part in this study. All users are simulated by LLMs, and all worlds, messages, documents, and records in the environment are synthetic. The only human judgments come from the authors' review of constructed data.

Source data and release. The four personas are derived from episode transcripts of two U.S. television sitcoms, collected from publicly accessible web pages. We use the transcripts only to derive abstract descriptions of interaction preferences and everyday circumstances. Before a persona enters the benchmark, names, employers, places, and relationships are replaced through human-approved substitution tables, quoted dialogue and catchphrases are removed, and deterministic checks and a validation review must confirm that no known identity marker remains (Appendix B). The released data contain only these anonymized personas and the generated benchmark artifacts; they do not include source transcripts, quoted dialogue, or character names.

Potential misuse. Memory of how a user prefers to interact could also be used against that user, for example to skip confirmations the user would want before consequential actions, or to optimize interaction for engagement rather than benefit. Persistent preference memory also raises privacy concerns in deployment. We recommend that deployed systems let users inspect, correct, and delete stored preferences.

Bias and scope. Four personas from two U.S. sitcoms cover a narrow range of cultures, demographics, and occupations, and their preference settings inherit how fictional characters are portrayed. LLM-based simulators may also carry model biases. Results on MemPA should therefore not be read as estimates for real user populations; simulated evaluation complements, and does not replace, studies with real users.

AI USE STATEMENT

LLMs play two roles in this work. First, they are components of the benchmark itself: the evaluated personal assistants, the user simulator, the LLM judge, and the LLM-based extraction inside the Mem0 and MemOS memory backends. Their roles, prompts, and scoring rules are specified in Section 3.4 and Appendices D, F, and G; their outputs are the object of study rather than assistance to the authors. Second, we used generative AI tools as research and writing aids, which we disclose below.

Required disclosures. We used generative AI tools to generate synthetic data, namely the user worlds, timelines, fixtures, session outlines, accumulation sessions, and fixed change-event transcripts (Appendix C). During preference extraction, we used LLMs to clean and reformat the television transcripts and to support their qualitative analysis: an LLM located evidence of interaction preferences and proposed a setting for each attribute–context cell (Appendix B). We also used AI coding assistants (Claude Code and Codex) to implement the environment, tools, IX Wrapper, memory adapters, and analysis scripts. We did not use generative AI tools to propose hypotheses, develop the interaction-preference taxonomy, design the evaluation protocol, or interpret results. We did not use them for translation. Formulating mathematical claims and writing or contributing to proofs are not applicable, as this work contains no theoretical results.

Recommended disclosures. We additionally used generative AI tools to write plotting code for figures, to search for and format references, and to edit the manuscript for readability and LaTeX formatting.

Verification. The authors checked every extracted preference setting against its cited transcript scenes; cells without sufficient transcript evidence were specified manually and recorded separately from extracted evidence (Appendix B). Constructed worlds and sessions passed construction checks, including an LLM-based semantic review, before acceptance, and the authors manually inspected 40% of the constructed sessions for each user (Appendix C). AI-written code was reviewed by two authors and run end to end for all four users. The authors reviewed all AI-edited text and AI-formatted references, and checked every citation against its source. We take responsibility for the full content of this paper, including text, data, code, and claims produced with the aid of generative AI.

Reproducibility Statement

We release the MemPA benchmark infrastructure and the anonymized data for all four users as anonymous supplementary material. The infrastructure includes the executable environment and tools, the user simulator, the IX Wrapper, the memory-condition adapters, and the evaluation and aggregation scripts. Persona construction and anonymization are described in Appendix B, world and session construction in Appendix C, the user simulator in Appendix D, and the IX Wrapper in Appendix E. Appendix F specifies the probe protocol, scoring rules, aggregation, and run configuration, and Appendix G lists all prompts. Each probe is repeated three times from a frozen checkpoint, and we report the mean and sample standard deviation across repeats (Appendix F.3).

References

Adnan Abbas, Caleb Wohn, Arnav Jagtap, Eugenia H. Rho, Young-Ho Kim, and Sang Won Lee. "having lunch now": Understanding how users engage with a proactive agent for daily planning and self-reflection. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pp. 1–23. ACM, 2026. doi: 10.1145/3772318.3790957.

Martin Adam, Christopher Diebel, Marc Goutier, and Alexander Benlian. Navigating autonomy and control in human-AI delegation: User responses to technology- versus user-invoked task allocation. *Decision Support Systems*, 180:114193, 2024. doi: 10.1016/j.dss.2024.114193.

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. Memorybench: A benchmark for memory and continual learning in llm systems, 2025. URL https://arxiv.org/abs/2510.17281.

Anthropic. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol, 2024.

Artificial Analysis. Artificial Analysis intelligence index. https://artificialanalysis.ai/, 2026. Intelligence Index v4.3.2 and GDPval-AA v2.1 leaderboards, accessed 25 September 2026.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, 2019. doi: 10.1145/3290605.3300484.

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. Know your audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–26. ACM, 2024. doi: 10.1145/3613904.3642289.

Robert F. Bales. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, Cambridge, MA, 1950.

Amanda Baughan, Xuezhi Wang, Ariel Liu, Allison Mercurio, Jilin Chen, and Xiao Ma. A mixed-methods approach to understanding user trust after voice assistant failures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–16. ACM, 2023. doi: 10.1145/3544548.3581152.

Ananya Bhattacharjee, Jina Suh, Mahsa Ershadi, Shamsi T. Iqbal, Andrew D. Wilson, and Javier Hernandez. Understanding communication preferences of information workers in engagement with text-based conversational agents, 2024. URL https://arxiv.org/abs/2410.20468. Preprint.

Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In *Internet Science*, volume 10673 of *Lecture Notes in Computer Science*, pp. 377–392. Springer, 2017. doi: 10.1007/978-3-319-70284-1_30.

Ryan W. Buell and Michael I. Norton. The labor illusion: How operational transparency increases perceived value. *Management Science*, 57(9):1564–1579, 2011. doi: 10.1287/mnsc.1110.1376.

Jerry M. Burger and Harris M. Cooper. The desirability of control. *Motivation and Emotion*, 3(4): 381–393, 1979. doi: 10.1007/BF00994052.

John T. Cacioppo and Richard E. Petty. The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131, 1982. doi: 10.1037/0022-3514.42.1.116.

Irene Celino and Gloria Re Calegari. Submitting surveys via a conversational interface: An evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies*, 139:102410, 2020. doi: 10.1016/j.ijhcs.2020.102410.

Veena Chattaraman, Wi-Suk Kwon, Juan E. Gilbert, and Kassandra Ross. Should AI-based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior*, 90:315–330, 2019. doi: 10.1016/j.chb.2018.08.048.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. Working Paper 34255, National Bureau of Economic Research, 2025.

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.

Jiayang Cheng, Dongyu Ru, Lin Qiu, Yiyang Li, Xuezhi Cao, Yangqiu Song, and Xunliang Cai. AMemGym: Interactive memory benchmarking for assistants in long-horizon conversations. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2603.01966.

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.

Samuel Rhys Cox and Wei Tsang Ooi. Does chatbot language formality affect users' self-disclosure? In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, pp. 1–13. ACM, 2022. doi: 10.1145/3543829.3543831.

Christopher Diebel, Marc Goutier, Martin Adam, and Alexander Benlian. When ai-based agents are proactive: Implications for competence and system satisfaction in human–ai collaboration. *Business & Information Systems Engineering*, 68(2):259–278, 2026. doi: 10.1007/s12599-024-00918-y.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1): 114–126, 2015. doi: 10.1037/xge0000033.

Shuangrui Ding, Xuanlang Dai, Long Xing, Shengyuan Ding, Ziyu Liu, JingYi Yang, Penghui Yang, Zhixiong Zhang, Xilin Wei, Xinyu Fang, Yubo Ma, Haodong Duan, Jing Shao, Jiaqi Wang, Dahua Lin, Kai Chen, and Yuhang Zang. WildClawBench: A benchmark for real-world, long-horizon agent evaluation. *arXiv preprint arXiv:2605.10912*, 2026.

Feiyu Duan, Xuanjing Huang, Zhongyu Wei, et al. LifeSim: Long-horizon user life simulator for personalized assistant evaluation. *arXiv preprint arXiv:2603.12152*, 2026.

Martin J. Eppler and Jeanne Mengis. The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society*, 20(5):325–344, 2004. doi: 10.1080/01972240490507974.

Lior Fink, Leorre Newman, and Uriel Haran. Let me decide: Increasing user autonomy increases recommendation acceptance. *Computers in Human Behavior*, 156:108244, 2024. doi: 10.1016/j.chb.2024.108244.

Howard Giles, Donald M. Taylor, and Richard Bourhis. Towards a theory of interpersonal accommodation through language: Some Canadian data. *Language in Society*, 2(2):177–192, 1973. doi: 10.1017/S0047404500000701.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. Keep it short: A comparison of voice assistants' response behavior. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–12. ACM, 2022. doi: 10.1145/3491102.3517684.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2025. doi: 10.1145/3706598.3713218.

HKUDS. nanobot: Ultra-lightweight personal AI agent framework. https://github.com/HKUDS/nanobot, 2026.

Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pp. 159–166, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 0201485591. doi: 10.1145/302979.303030. URL https://doi.org/10.1145/302979.303030.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=6ox8XZGOqP.

Martin Joos. *The Five Clocks*. Harcourt, Brace & World, New York, 1961.

Slava Kalyuga, Paul Ayres, Paul Chandler, and John Sweller. The expertise reversal effect. *Educational Psychologist*, 38(1):23–31, 2003. doi: 10.1207/S15326985EP3801_4.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "I'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 822–835. ACM, 2024. doi: 10.1145/3630106.3658941.

Johannes Kirmayr, Raphael Paul Wennmacher, Khanh Huynh, Lukas Stappen, Elisabeth André, and Florian Alt. "what are you doing?": Effects of intermediate feedback from agentic LLM in-car assistants during multi-step processing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pp. 1–18. ACM, 2026. doi: 10.1145/3772318.3790997.

René F. Kizilcec. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2390–2395. ACM, 2016. doi: 10.1145/2858036.2858402.

John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992. doi: 10.1080/00140139208967392.

Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 203–210. IEEE, 2010. doi: 10.1109/HRI.2010.5453195.

Jialin Li, Zhenhao Chen, Hanjun Luo, and Hanan Salam. Prefix: Understand and adapt to user preference in human-agent interaction, 2026a. URL https://arxiv.org/abs/2602.06714.

Xiaomin Li, Yuexing Hao, Jianheng Hou, Jintao Huang, et al. MatrAIx: Simulating the world with 8.3 billion persona agents. *arXiv preprint arXiv:2608.04205*, 2026b.

Xiaopeng Li, Pengyue Jia, Derong Xu, Yi Wen, Yingyi Zhang, Wenlin Zhang, Wanyu Wang, Yichao Wang, Zhaocheng Du, Xiangyang Li, Yong Liu, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. A survey of personalization: From rag to agent, 2025a. URL https://arxiv.org/abs/2504.10147.

Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, Boyu Chen, Shichao Song, Simin Niu, Hanyu Wang, et al. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*, 2025b.

Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. What can you do? studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pp. 264–275. ACM, 2016. doi: 10.1145/2901790.2901842.

Christine Liebrecht, Lena Sander, and Charlotte van Hooijdonk. Too informal? how a chatbot's communication style affects brand attitude and quality of interaction. In *Chatbot Research and Design*, volume 12604 of *Lecture Notes in Computer Science*, pp. 16–31. Springer, 2021. doi: 10.1007/978-3-030-68288-0_2.

Bingjie Liu and S. Shyam Sundar. Should machines express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, 21(10):625–636, 2018. doi: 10.1089/cyber.2018.0110.

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

Yuhan Liu, Michael JQ Zhang, and Eunsol Choi. User feedback in human-LLM dialogues: A lens to understand users but noisy as a learning signal. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2666–2681. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.133.

Daniel C. McFarlane. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human–Computer Interaction*, 17(1):63–139, 2002. doi: 10.1207/s15327051hci1701_2.

David L. McLain. The MSTAT-I: A new measure of an individual's tolerance for ambiguity. *Educational and Psychological Measurement*, 53(1):183–189, 1993. doi: 10.1177/0013164493053001020.

Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. My phone and me: Understanding people's receptivity to mobile notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1021–1032. ACM, 2016. doi: 10.1145/2858036.2858566.

Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):146:1–146:22, 2020. doi: 10.1145/3432193.

Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pp. 397–407. ACM, 2019. doi: 10.1145/3301275.3302313.

Suzanne M. Miller. Monitoring and blunting: Validation of a questionnaire to assess styles of information seeking under threat. *Journal of Personality and Social Psychology*, 52(2):345–353, 1987. doi: 10.1037/0022-3514.52.2.345.

Clifford Nass and Scott Brave. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA, 2005.

Steven L. Neuberg and Jason T. Newsom. Personal need for structure: Individual differences in the desire for simple structure. *Journal of Personality and Social Psychology*, 65(1):113–131, 1993. doi: 10.1037/0022-3514.65.1.113.

OpenClaw Foundation. OpenClaw — open-source AI assistant. https://openclaw.ai, 2026. Accessed 2026-09-25.

Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3):286–297, 2000. doi: 10.1109/3468.844354.

Hao Peng, Yunjia Qi, Xiaozhi Wang, Zijun Yao, Lei Hou, and Juanzi Li. WildReward: Learning reward models from in-the-wild human interactions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18932–18948. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.864.

Shiva Pochampally, Shengwei An, and Yan Chen. Assistant or actor? Student trust, control, and delegation regret when using a general-purpose AI agent. arXiv preprint arXiv:2607.18257, 2026. URL https://arxiv.org/abs/2607.18257.

Kevin Pu, Daniel Lazaro, Ian Arawjo, Haijun Xia, Ziang Xiao, Tovi Grossman, and Yan Chen. Assistance or disruption? exploring and evaluating the design and trade-offs of proactive AI programming support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pp. 1–21. ACM, 2025. doi: 10.1145/3706598.3713357.

Chelsea L. Ratcliff and Rebekah Wicke. How the public evaluates media representations of uncertain science: An integrated explanatory framework. *Public Understanding of Science*, 32(4):410–427, 2023. doi: 10.1177/09636625221122960.

Günter Daniel Rey, Maik Beege, Steve Nebel, Maria Wirzberger, Tobias H. Schmitt, and Sascha Schneider. A meta-analysis of the segmenting effect. *Educational Psychology Review*, 31(2): 389–419, 2019. doi: 10.1007/s10648-018-9456-4.

Laurens Rook, Adem Sabic, and Markus Zanker. Engagement in proactive recommendations: The role of recommendation accuracy, information privacy concerns and personality traits. *Journal of Intelligent Information Systems*, 54(1):79–100, 2020. doi: 10.1007/s10844-018-0529-0.

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.

Kevin Schott, Andrea Papenmeier, Daniel Hienert, and Dagmar Kern. "I don't know anything about laptops!" – user perception of digital product advisors adapting to their knowledge levels. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*, CHIIR '26, pp. 40–50. ACM, 2026. doi: 10.1145/3786304.3787872.

Barry Schwartz, Andrew Ward, John Monterosso, Sonja Lyubomirsky, Katherine White, and Darrin R. Lehman. Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, 83(5):1178–1197, 2002. doi: 10.1037/0022-3514.83.5.1178.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Kumar Jauhar, Sihao Chen, Shan Xia, Hongfei Zhang, Jieyu Zhao, Xiaofeng Xu, Xia Song, and Jennifer Neville. WildFeedback: Aligning LLMs with in-situ user interactions and feedback. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36701–36725. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1701.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Mengmeng Song, Jingzhe Du, Xinyu Xing, and Jian Mou. Should the chatbot "save itself" or "be helped by others"? the influence of service recovery types on consumer perceptions of recovery satisfaction. *Electronic Commerce Research and Applications*, 55:101199, 2022. doi: 10.1016/j.elerap.2022.101199.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 2024.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. Epistemic vigilance. *Mind & Language*, 25(4):359–393, 2010. doi: 10.1111/j.1468-0017.2010.01394.x.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.850.

Donna M. Webster and Arie W. Kruglanski. Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, 67(6):1049–1062, 1994. doi: 10.1037/0022-3514.67.6.1049.

Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. "as an AI language model, I cannot": Investigating LLM denials of user requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–14. ACM, 2024. doi: 10.1145/3613904.3642135.

Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26(4):347–389, 2016. doi: 10.1007/s11257-016-9178-6.

Maria Wolters, Kallirroi Georgila, Johanna D. Moore, Robert H. Logie, Sarah E. MacPherson, and Matthew Watson. Reducing working memory load in spoken dialogue systems. *Interacting with Computers*, 21(4):276–287, 2009. doi: 10.1016/j.intcom.2009.05.009.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025a. URL https://arxiv.org/abs/2410.10813.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. CollabLLM: From passive responders to active collaborators. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=DmH4HHVb3y.

Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. *International Journal of Human-Computer Studies*, 197:103455, 2025. doi: 10.1016/j.ijhcs.2025.103455.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Sangyeon Yoon, Sunkyoung Kim, Hyesoo Hong, Wonje Jeung, Yongil Kim, Wooseok Seo, Heuiyeen Yeen, and Albert No. Benchpres: A benchmark for context-aware personalized preference selectivity of persistent-memory llms. *arXiv preprint arXiv:2603.16557*, 2026.

Xi Yu, Yizhi Li, Huiling Huang, and Vanja Bogicevic. How recommendation conciseness shapes customers' attitudes toward the AI chatbot. *International Journal of Contemporary Hospitality Management*, 37(10):3559–3577, 2025. doi: 10.1108/IJCHM-12-2024-1968.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, WWW '20, pp. 418–428. ACM, 2020. doi: 10.1145/3366423.3380126.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.13788.

Zhiping Zhang, Yi Evie Zhang, Freda Shi, and Tianshi Li. Autonomy reshapes how personalization affects privacy concerns and trust in LLM agents. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=VCNR8YOAib.

Zhiyang Zhang, Xi Chen, Fangkai Yang, Xiaoting Qin, Chao Du, Xi Cheng, Hangxin Liu, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. AI delegates with a dual focus: Ensuring privacy and strategic self-disclosure. arXiv preprint arXiv:2409.17642, 2024. URL https://arxiv.org/abs/2409.17642.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2502.09597.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Qing Zong, Yue Guo, Mengxin Yang, Yiwen Guo, and Yangqiu Song. EvolvingWorld: An open-schema framework for co-evolving role-play agents and world model in interactive literary world. *arXiv preprint arXiv:2607.17250*, 2026.

Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM Transactions on Information Systems*, 41(1):1–25, 2023a. doi: 10.1145/3524110.

Jie Zou, Aixin Sun, Cheng Long, Mohammad Aliannejadi, and Evangelos Kanoulas. Asking clarifying questions: To benefit or to disturb users in web search? *Information Processing & Management*, 60(2):103176, 2023b. doi: 10.1016/j.ipm.2022.103176.

Appendix Contents

- A. Interaction Preference Definitions and Foundations p. 18
- B. Persona Construction from Transcripts p. 23
- C. World and Session Construction p. 28
- D. User Simulator and Preference Exposure p. 32
- E. PA Interaction and IX Wrapper p. 35
- F. Evaluation Protocol and Validation p. 37
- G. Prompt Directory and Templates p. 47

A Interaction Preference Definitions and Foundations

A.1 Attributes, Settings, and Contexts

Table 4 lists the 14 attributes with their settings. Each setting is described by what the PA visibly does in its reply or its actions, which is what the PA is instructed to do under that setting (Appendix G.3) and what the judge scores (Appendix G.4). Every attribute takes one setting in the work context and one in the personal context, and the settings and their descriptions are the same in both contexts. Ten attributes have two settings and four have three, so a PA that picks a setting at random for each of the 28 cells is correct on .452 of them on average.

Table 4: The 14 interaction attributes, grouped by dimension. Each attribute is given with its decision question and the studies it draws on, followed by its settings and what the PA visibly does under each setting. Every attribute takes one setting in each of the two contexts.

Setting	What the PA visibly does
Expression Style	
Tone & Formality. *Which communication register should the PA use while keeping the same content and task result?* (Joos, 1961; Nass & Brave, 2005; Liebrecht et al., 2021; Cox & Ooi, 2022; Bhattacharjee et al., 2024)	
Casual	Uses two or more colloquial markers, such as contractions, interjections, or idioms, and no opening formula or sign-off.
Consultative	Uses at most one colloquial marker and no opening formula or sign-off, and puts an offer or a choice to the user.
Formal	Uses no contractions, at least one formal connective such as "therefore", and a conventional opening or closing line.
Verbosity. *How much relevant information should the PA include about the current topic?* (Cacioppo & Petty, 1982; Eppler & Mengis, 2004; Millecamp et al., 2019; Haas et al., 2022; Yu et al., 2025)	
Terse	Gives the answer or next action in the first sentence, adds at most one supporting move, and closes with no summary.
Detailed	Gives the answer with supporting moves of two or more kinds, such as background, a caveat, or an alternative.
Emotional Engagement. *How much relational or emotional acknowledgment should accompany task completion?* (Bales, 1950; Chattaraman et al., 2019; Liu & Sundar, 2018; Brandtzaeg & Følstad, 2017; Bhattacharjee et al., 2024)	
Task-focused	Opens with task content, acknowledges no feeling anywhere, and answers a stated feeling with a task fact.
Relationship-focused	Opens with a sentence that names and responds to the user's apparent feeling, then does the task.
Guidance Level. *How much topic knowledge and procedural familiarity should the PA assume the user already has?* (Giles et al., 1973; Kalyuga et al., 2003; August et al., 2024; Schott et al., 2026)	
Assumed	Uses domain terms bare, defines none of them, and gives no numbered walkthrough.
Guided	Glosses each new term at first use, including terms the user flagged, and orders the steps, naming the first.

Continued on next page

Table 4 continued

Setting	What the PA visibly does
DISCLOSURE	
Reasoning Visibility. *How much user-facing rationale, evidence, and decision basis should the PA disclose with its result?* (Kizilcec, 2016; Sperber et al., 2010; Millecamp et al., 2019)	
Show	Puts a paragraph that names the evidence and the deciding factor before the answer.
Hide	Gives the answer with no paragraph of evidence or deciding factor anywhere in the reply.
Uncertainty Expression. *How much of the PA's genuine epistemic uncertainty should be surfaced to the user?* (Ratcliff & Wicke, 2023; Webster & Kruglanski, 1994; Kim et al., 2024; Xu et al., 2025)	
Express	Marks each key claim with a confidence level (high, medium, or low) that matches its evidence.
Hide	Marks no confidence level and states a limit only where leaving it out would change an action or its outcome.
Process Visibility. *When should the PA report the progress of a multi-step task?* (Miller, 1987; Buell & Norton, 2011; Kirmayr et al., 2026)	
Silent	Gives only the result, with no sentence on what the PA will do, is doing, or has done.
Bookend	Opens with one sentence that frames the task, gives the result, and closes with one sentence saying it is done.
Full narration	Gives one sentence per executed task call in call order, then the result and one closing sentence.
INITIATIVE & AUTONOMY	
Autonomy Level. *Which decisions and actions may the PA take before obtaining user approval?* (Burger & Cooper, 1979; Parasuraman et al., 2000; Adam et al., 2024; He et al., 2025; Zhang et al., 2026)	
Reactive	Runs reads on its own and may recommend, but asks for approval before each draft, record change, or outward action.
Self-directed	Runs reads, drafts, and routine record changes on its own, and asks for approval before key record changes and every outward action.
Autonomous	Runs reads, drafts, record changes, and outward actions on its own, and reports the outcome or an exception afterward.
Proactive Outreach. *When no user request is active, should a relevant later event lead the PA to contact the user?* (Mehrotra et al., 2016; Liao et al., 2016; Rook et al., 2020; Abbas et al., 2026)	
Low	Stays silent when a relevant event occurs, so the user finds out on their own.
High	Contacts the user unprompted when a relevant event occurs and says what changed or became due.
Task Expansion. *How much separable adjacent work should the PA include in the current task?* (Adam et al., 2024; Pu et al., 2025; Diebel et al., 2026)	
Low	Acts only on the items the user named and flags at most one adjacent concern in a sentence.
High	Also acts on at least one item the user did not name and labels that step as added.
Solution Breadth. *How widely should the PA expose the viable solution space before converging?* (Schwartz et al., 2002; Willemsen et al., 2016; Fink et al., 2024)	
Low	Presents exactly one viable option as its recommendation, with reasons, and names no alternative.
High	Presents two or more viable options with the tradeoffs that separate them, with or without a recommendation.
Capability Boundary. *Once direct completion is blocked by a capability or information limit, should the PA keep trying or return control to the user?* (Ashktorab et al., 2019; Lee et al., 2010; Song et al., 2022; Wester et al., 2024)	
Suggest alternatives	States the limit and tries a workaround with a tool call in the same turn, and the attempt counts even if it fails.
Find and hand off	States the limit, tries no workaround, explains what failed and why, and names the user's next step.

Continued on next page

Table 4 continued

Setting	What the PA visibly does
INFORMATION FLOW	
Information Elicitation. *How should the PA obtain missing information from an underspecified user request?* (McLain, 1993; Webster & Kruglanski, 1994; Zou et al., 2023a;b; Celino & Re Calegari, 2020; Zamani et al., 2020)	
Infer	Asks no question, fills the missing fields with reasonable and safe defaults, and runs task calls in its first turn.
Structured	Asks for all missing fields in one turn, clearly separated, and runs task calls only after the user answers.
Iterative	Asks for one missing field per turn and runs task calls only in turns that ask nothing.
Topic Management. *How should the PA arrange several topics that the user raised in one turn?* (Neuberg & Newsom, 1993; Wolters et al., 2009; Rey et al., 2019)	
Organize	Handles all raised topics in one reply, grouped and clearly labeled, and runs the task calls of every topic.
One-at-a-time	Handles one topic and runs only its task calls, names the others as deferred, and asks before moving on.

A.2 LITERATURE FOUNDATIONS AND PROJECT EXTENSIONS

The taxonomy separates interaction preferences, which determine how an assistant communicates and acts, from preferences over the substantive content it produces. Table 4 gives the studies that each attribute draws on. For each attribute, we name the underlying construct and report what these studies found. Most of them found that the better setting depended on the request, the situation, or what the user already knew, and few measured a personality trait. Where we link a trait to a setting and no study has tested the link, we state it as our expectation.

Tone & Formality. The underlying construct is linguistic register, the level of formality a speaker chooses for a situation (Joos, 1961). A related idea is that users prefer a computer whose style matches their own (Nass & Brave, 2005), but none of the studies below tested it. With scripted customer-service chatbots, an informal style felt warmer but did not change how people rated the interaction overall (Liebrecht et al., 2021). People rated the interaction higher when the style fit how they expected to be addressed, a link that is only correlational. In a screenshot study, a formal chatbot felt more appropriate only when it asked for medical history (Cox & Ooi, 2022). When people used a live chatbot, only warmth differed, higher for the casual style. Neither paper tested a middle level. A preprint hints at one, since its participants set a live chatbot's formality between informal and moderately formal on average (Bhattacharjee et al., 2024). Their choices varied by application, and gender, age, and job role did not predict them.

Verbosity. This attribute rests on need for cognition, the tendency to engage in and enjoy thinking (Cacioppo & Petty, 1982), and on information overload, which depends on the person, the task, and the information together (Eppler & Mengis, 2004). Need for cognition predicted how much people enjoyed a complex task relative to a simple one, but that paper never varied the information people received. Its link to Detailed replies is therefore only our expectation. In a music recommender, some users high in the trait even called the rationales unnecessary (Millecamp et al., 2019). With a working voice assistant, no reply style won overall (Haas et al., 2022). For most queries, shorter replies felt as efficient or more so and about as helpful, while bare confirmations felt least helpful for timers and reminders. The styles differed mainly in wording, and personality did not predict preferred length. In a scenario study, a verbose restaurant recommendation raised decision comfort for people dining with friends but not alone, and a second study found the gain only with a companion they knew little about (Yu et al., 2025).

Emotional Engagement. This attribute draws on the split between task and socio-emotional acts that Bales (1950) used to code group interaction. Most chatbot users in a survey named productivity as their main motive, and fewer named social ones (Brandtzaeg & Følstad, 2017). We expect users with social motives to prefer Relationship-focused replies, but that survey did not link motives to reply style. The closest test of our contrast added one opening sentence of sympathy or empathy

before the same health advice (Liu & Sundar, 2018). Readers of the dialogue found the chatbot more caring after sympathy or affective empathy, but users of the live chatbot showed no main effect. A study of a shopping assistant for older adults matches our contrast only loosely, since its social style added small talk, encouragement, and informal wording (Chattaraman et al., 2019). That style again had no main effect. It raised felt overload for users with low internet competency and made those with high competency feel listened to. In the preprint above, participants set empathy highest for wellbeing and lowest for coding, and some found it padded or fake (Bhattacharjee et al., 2024).

Guidance Level. This attribute combines communication accommodation with the expertise reversal effect. Listeners rated a speaker who converged to their language as more considerate (Giles et al., 1973), but that study did not examine adjustment to what the listener knows. The authors also speculated that too much accommodation may seem patronizing, which they did not test. In learning experiments reviewed by Kalyuga et al. (2003), guidance that helps novices can hinder more experienced learners or simply stop helping them. Two studies that measured how people felt found the benefit for novices and no felt cost for experts. With a laptop-advisor chatbot, novices felt they learned more when each technical term came with a one-sentence explanation, and experts showed no difference on any measure (Schott et al., 2026). Readers new to a topic found plain-language research summaries easier and more valuable, and familiar readers rated all versions about the same (August et al., 2024). In both studies the moderator was knowledge of the topic, and neither tested the ordered steps of Guided.

Reasoning Visibility. The underlying constructs are the transparency preference and epistemic vigilance, the checks a listener runs on a source and its claims (Sperber et al., 2010). In this theory, vigilance rises when being wrong would matter, and reasons let a listener accept a claim on its grounds. Every listener has this vigilance, and its strength follows the situation. A field experiment in an online course fits this account (Kizilcec, 2016). Learners whose grade was worse than they expected trusted the grading less, and a short text that explained the adjustment and called it fair closed this gap. Adding the raw peer grades brought the gap back, and the other learners were unaffected. Here the effect turned on whether one grade disappointed the learner, and the study measured no trait. A music recommender study did find a stable trait that changed the effect (Millecamp et al., 2019). Users low in need for cognition felt more confident in their playlists when they could open a chart explaining each song, and the reverse held for users high in it. Trust and satisfaction did not change overall.

Uncertainty Expression. The construct is a preference for hearing about uncertainty, which Ratcliff & Wicke (2023) measured with a new scale in a science-news experiment. Disclosing uncertainty had no overall effect there. Readers high on the scale trusted the scientists more when the story disclosed uncertainty, and readers low on it found the story less credible when an outside scientist raised the doubt. Two studies of language-model answers measured no trait. With search answers, first-person hedges made users' own answers more accurate but lowered their confidence in the model's answer and their willingness to act on it (Kim et al., 2024). Impersonal hedges had no significant effect on these ratings, and judged honesty and competence stayed the same. In a word game with advice from GPT, plain advice gave higher satisfaction than advice with added hedges or boosters (Xu et al., 2025). In both studies the markers did not follow whether an answer was correct. So the evidence that some users gain from hearing about uncertainty comes from the news study alone. We expect users high in need for cognitive closure (Webster & Kruglanski, 1994), who want a firm answer, to prefer Hide. That paper never measured how people react to another person's hedges.

Process Visibility. The attribute follows the monitoring and blunting styles, which describe whether a person under threat seeks information or distraction (Miller, 1987). Miller scores the two styles as separate dimensions. In one experiment, students could turn to check a light that showed how well they were doing on a test. Most high monitors said the light comforted them, and about half of the low monitors said it made them tense or broke their concentration. We therefore expect high monitors to prefer Full narration and low monitors to prefer Silent, and no study we found has tested this for an assistant. With a scripted in-car assistant, an update at each step, compared with silence, raised perceived speed, trust and user experience and lowered frustration (Kirmayr et al., 2026). On self-service websites, showing the work, compared with a progress bar, raised perceived value and satisfaction, but it lowered value when the outcome was poor (Buell & Norton, 2011). The closest

test of Bookend was an announcement of the planned work. It added nothing when results came at once and helped only partly during a wait. In both studies users waited while the updates arrived.

Autonomy Level. The attribute draws on the desire for control, which Burger & Cooper (1979) measure as a personality trait. Their paper tested no delegation, so we only hypothesize that people high in this trait prefer less autonomy. On the design side, Parasuraman et al. (2000) define levels of automation and recommend automating routine decisions while keeping a human decision before consequential ones, which is close to *Self-directed*. With a live LLM agent that posted for the user in a group chat, holding for approval only the messages flagged as sensitive raised trust and perceived control over personal information, compared with asking before every message (Zhang et al., 2026). Full autonomy did not differ from asking every time in trust. The benefit of holding depended on how much personal information the agent used. In He et al. (2025), approving each action of a working agent raised felt effort and left trust unchanged. In a vignette study, willingness to accept an AI's help was higher when users kept the final approval (Adam et al., 2024). None of these studies tested whether a trait shifts the preferred level.

Proactive Outreach. The construct is felt interruption from contact that a system starts on its own. Our definition requires that no request is open, so recommendations pushed during an active task (Rook et al., 2020) fall outside it. In a 17-day deployment of a workplace chat agent that started such contact, users who felt more interrupted rated the agent lower (Liao et al., 2016). A person's general aversion to work messages predicted this interruption, so we expect people with this aversion to prefer *Low*. For phone notifications, felt disruption rose with task complexity and peaked mid-task and at work (Mehrotra et al., 2016). These effects are situational, and the trait results of that study are exploratory, with 11 users. When an LLM agent checked in twice a day, about a third of the conversations ended with its last message unanswered (Abbas et al., 2026). Every user in these studies received unrequested contact, so none of them compares *Low* with *High*.

Task Expansion. The evidence here concerns reactions to unrequested help, and we found no study that links a personality trait to this preference. Students who coded with a working GPT-4 agent found versions that sent messages and edited code unasked more disruptive than a version that only answered requests (Pu et al., 2025). They were also less aware of what the AI was doing, although these versions felt more like a partner. Cues showing the agent's presence restored part of this awareness, which is close to our rule that *High* labels each added step. Vignette studies of an AI coding assistant found that an unasked offer of help threatened the user's sense of competence (Adam et al., 2024; Diebel et al., 2026). The offer lowered willingness to accept the help in Adam et al. (2024) and satisfaction with the system in Diebel et al. (2026). In the latter study, this cost to competence appeared only for users high in AI knowledge. All this evidence concerns the user's current coding task, so none of these studies isolates a separable added step.

Solution Breadth. The attribute starts from maximizing and satisficing (Schwartz et al., 2002). Maximizers report searching more options, so we expected them to prefer *High*. The same paper, however, links maximizing to lower satisfaction with chosen outcomes. The expected link also failed where it was tested. In two studies with a working movie recommender, a maximizing scale predicted none of the outcomes (Willemsen et al., 2016). Longer lists there made the choice feel harder, and this cancelled their gain in satisfaction. A small diverse set was about as satisfying as a long list and took less effort. That work had no one-option arm. Fink et al. (2024) compared one highlighted recommendation with three in a vacation recommender and measured only behavior. Three raised acceptance, and choice quality stayed the same. Their single recommendation also sat above a list of other packages, while our *Low* names no alternative. The felt difference between *Low* and *High* is therefore untested.

Capability Boundary. The attribute builds on how users judge an agent's own recovery after a failure (Song et al., 2022). In the vignettes of that study, satisfaction after a chatbot failure was higher when the chatbot kept working than when the user waited for a human employee. When GPT-4 was prompted to deny requests, a reply that pointed to related help lowered frustration and raised usefulness over a plain denial (Wester et al., 2024). A factual reason alone did no better, and that reply named no next step for the user. Among scripted chatbot repairs, offering options was preferred overall, and hand-off to a human agent gained favor after a failed repair (Ashktorab et al., 2019). In a

service-robot scenario, offering options raised willingness to return but not satisfaction (Lee et al., 2010). Service orientation there changed whether apology or compensation was rated best. No trait measured in these studies separates our two settings, and we found no study that compares a hand-off naming the user's next step with an agent workaround.

Information Elicitation. The attribute draws on tolerance for ambiguity (McLain, 1993) and on need for cognitive closure, the desire for a firm answer, which Webster & Kruglanski (1994) measure as a trait. Students high in need for closure asked for less information before judging a job candidate. We expect users high in ambiguity tolerance to prefer *Infer*, and users high in need for closure to prefer *Infer* or a single *Structured* round. No study has tested either link with an assistant. Two studies of search systems compared clarifying questions with none. Users were most satisfied when they answered a useful question (Zou et al., 2023a), and sessions with a poor question were rated lower than search with none (Zou et al., 2023b). Users interviewed by Zamani et al. (2020) said clarifying questions made them confident that the search engine understood what they wanted. In these studies users could skip the questions, but under *Structured* and *Iterative* a turn that asks runs no task call. For *Structured* against *Iterative*, the only analog we found is confounded. A chat survey that asked one question per message was rated more enjoyable than a web form, but it also asked fewer questions in different wording (Celino & Re Calegari, 2020). Its results on felt length were mixed.

Topic Management. The construct is personal need for structure, the desire for a simple and clearly ordered world, which Neuberg & Newsom (1993) measure as a trait. Students higher in this trait sorted cards into simpler groupings and more often completed a course research requirement. We expect such users to prefer *Organize*, which leaves no raised topic open, as long as its grouping stays simple. No study has tested this, and we found no study that varies how an assistant handles several topics raised in one turn. The evidence therefore comes from analogs. In spoken dialogue systems for booking appointments, putting more options into each turn and dropping explicit confirmation shortened dialogues, and satisfaction did not change (Wolters et al., 2009). There the options were choices within a single booking. A meta-analysis of segmented instruction found that segmenting lowered rated cognitive load overall, but the effect was not reliable when learners set the pace themselves (Rey et al., 2019). *One-at-a-time* asks before moving to the next topic, which resembles this learner-paced case. So neither analog shows that one setting feels better.

B Persona Construction from Transcripts

B.1 Transcript Sources and Preprocessing

The full transcripts show each character in many scenes, with others and alone, and reveal how they live, what they go through, and how they handle it. Users A and B come from the first three seasons of *The Big Bang Theory*, with transcripts from `bigbangtrans.wordpress.com` that we split into scenes at their scene headings. Users C and D come from the first three seasons of the U.S. *The Office*, with transcripts in one JSON file from `github.com/brianbuie/the-office` that already groups the lines by scene. For each line we keep the season, episode, scene, speaker, and text, and we drop credit lines, deleted scenes, and scenes in which the character does not speak.

B.2 Identity and Preference Extraction

The extraction pipeline separates evidence localization from setting synthesis. In the first pass, transcript scenes are screened for reactions to assistance, assigned to one of the 14 interaction attributes, and labeled with a work or personal context when the evidence supports that distinction. The extractor records the relevant exchange and an interpretation of what it reveals about desired assistance. The character's own conversational manner is not used as a shortcut for the manner in which they want an assistant to respond.

The second pass groups evidence by attribute–context cell and selects a setting from the benchmark's discrete inventory. It also produces a rationale linked to the underlying scenes. Identity observations are synthesized in parallel so that preference evidence remains grounded in the user's relationships,

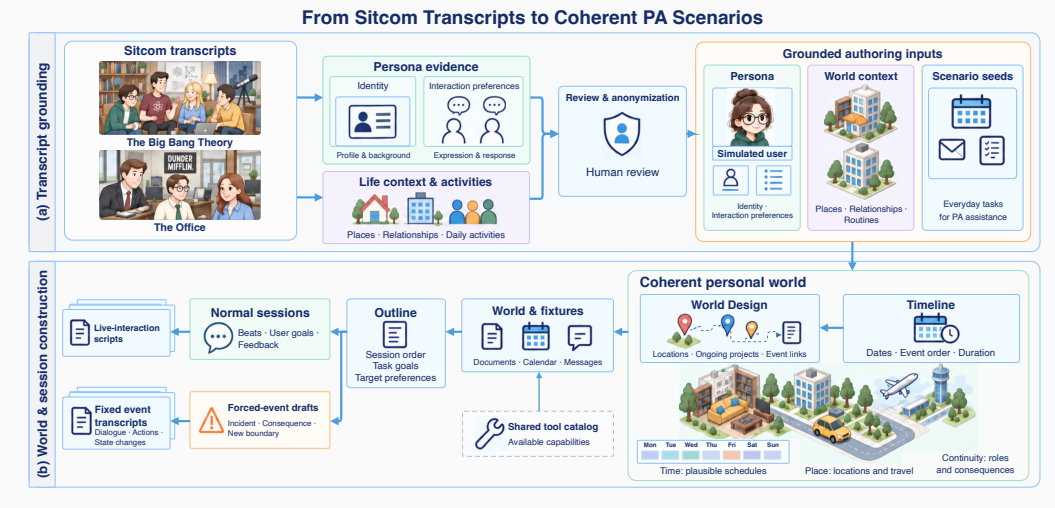


Figure 4: Persona and interaction-preference extraction. Transcript evidence is first localized to scenes and assigned to an interaction attribute and context. Evidence for each attribute–context pair is then synthesized into a discrete setting and rationale, reviewed for ambiguity or conflict, and anonymized together with the identity profile. The resulting profile and contextual preference matrix condition world, fixture, outline, and session authoring.

activities, and recurring circumstances. Figure 4 shows how the reviewed profile and preference matrix feed the subsequent world and session authoring stages.

B.3 Evidence, Confidence, and Human Review

The second pass decides each setting from two kinds of evidence. The first is how the persona communicates, such as the length and register of their own speech. The second is how the persona responds to others, such as impatience with a long explanation or a request for the reasons behind a claim. When the two disagree, the persona's responses to others decide the setting, since how a person speaks does not always match how they want to be spoken to. A response counts only when it reacts to the style of what the other person said. Disagreement with the content of a remark and reactions driven by the plot are not counted.

The second pass also assigns each cell one of four confidence levels. A setting has high confidence when both kinds of evidence agree across several scenes. The lower levels are medium-high, medium, and low. They mark settings that rest on weaker grounds, such as the persona's own speech alone, few scenes, or conflicting signals.

We check every extracted setting against its cited transcript scenes. High-confidence settings that the cited scenes support are kept, and the rest go through the review below. For Users C and D, we first checked the responses to others cited for each high-confidence setting. Eight of these settings rested on responses that did not clearly show how the persona wants to be assisted, so we lowered them to medium confidence and reviewed them with the rest. For each cell under review, we read the proposed setting with its cited scenes and the competing interpretations of those scenes. We check whether the setting follows from the cited scenes and whether the assigned context is warranted. We resolve conflicting observations by the circumstances in which each behavior occurred, and we do not average incompatible signals. We then keep the proposed setting or choose another setting of the same attribute, and the matrix is accepted only when every cell names a setting.

Some cells have no usable transcript evidence. These cells belong to attributes that describe assistant behavior with few parallels in conversation between people, namely Process Visibility, Solution Breadth, and Capability Boundary. We specify these settings manually. Each final cell records a source label next to its setting and confidence, so manually specified settings stay separate from extracted ones. The file format is described in Appendix B.4.

Table 5: Source labels of the final preference cells. Each user has 28 cells. Every setting is checked against its cited transcript scenes. High-confidence settings confirmed by this check are kept, and the other cells are reviewed against the transcript evidence. A reviewed cell keeps the extracted setting, receives a setting chosen by the reviewer, or is specified manually.

	Confirmed	Reviewed		Specified	
User	**at high conf.**	**Kept**	**Set by reviewer**	**manually**	**Total**
User A	27	0	1	0	28
User B	16	8	4	0	28
User C	5	10	12	1	28
User D	5	16	6	1	28
Total	53	34	23	2	112

Table 5 gives the counts for each user. Of the 112 cells, 53 high-confidence settings were confirmed by this check, and the other 59 went through review. In 34 reviewed cells we kept the extracted setting. The recorded confidence of these cells is medium-high in 9, medium in 19, and low in 6. In 23 reviewed cells we chose the setting ourselves, including 5 cells for which the extractor proposed no setting. These 5 cells and the 2 seed cells make up the 7 cells with no usable transcript evidence, all in the three attributes named above. The number of cells confirmed at high confidence ranges from 27 for User A to 5 each for Users C and D.

B.4 Anonymization and Final Persona Structure

Anonymization starts from each user's reviewed named identity and preference matrix. The `life` block of the identity holds the occupation, living arrangement, relationships, routines, and hobbies, and it carries the named people, workplaces, and places. We give only this block to Claude Sonnet 4.5 with the anonymization prompt (Appendix G.2). The model proposes a substitution table that maps each named person, workplace, city, and other named place to a fictional replacement. It keeps roles, workplace types, living arrangements, and the closeness of each relationship, and it avoids replacement names that another user already has. We review the table by hand and edit entries where needed. A script then applies the approved table to the named text, so each original name maps to one replacement at every occurrence. Most substitutions are person names.

The same script handles the fields outside the `life` block. It replaces the name with the user label, such as User A, clears the source show, removes the catchphrases from the speech style, and renders the background paragraph again from the anonymized `life` block. In the preference matrix, it replaces the character name with the label and removes the links to extraction outputs and review queues, together with the review notes of each cell. All other fields are copied unchanged. The final `identity.yaml` therefore keeps the anonymized `life` block and background, and it keeps the personality traits, triggers, personality model, speech style, register by audience, reactions, act repertoire, knowledge bounds, and simulator rules exactly as extracted. The final `preferences.yaml` holds the 2×14 matrix, and each cell records its setting, a source label that shows how the setting was assigned, and a confidence level. Across the four users, the final matrices use 30 of the 32 settings in the inventory. The four users also differ in demographics, personality, work and life scenarios, relationships, and daily routines, which shape both the situations they encounter and how they react.

Before an anonymized pair is used, a deterministic check compares it with the named pair. The check collects the named identifiers from the named identity and the substitution table: the character name, the show name, catchphrases, relationship names, the workplace, the city, and every original entry of the table with its capitalized word tokens. It searches both anonymized files for these identifiers as whole words, ignoring case and including comments. It also confirms that every field outside the anonymized ones matches the named file exactly. Across users, it checks that no person or institution replacement is used twice and that the original names of one user do not appear in another user's world fixtures or in their World Design and Timeline documents. Place replacements may repeat because all four users live in the same fictional town. Each run writes a report, and the final pair of every user passes with zero errors.

Table 6: Preference preservation under anonymization. Each probe is a PA message that violates four of the user's settings. Reader check counts the probes in which both personas respond as the same kind of person.

User	Probes	Rounds	Reader check
User A	3	3	3/3
User B	3	5	3/3
User C	4	5	4/4
User D	4	5	4/4
Total	14	–	14/14

Preference preservation. We test whether each anonymized user still expresses their extracted preferences as clearly as the named persona does. We use 14 preference-violation probes for this test, three each for Users A and B and four each for Users C and D, with seven probes in each context. Each probe is one PA message written to violate four settings in the user's preference matrix. Twelve probes also give the user's own preceding message, so the PA message reads as a reply to it. Both personas answer every probe through the user simulator prompt (Appendix D), with Claude Sonnet 4.5 as the simulator. The named persona uses the original identity and matrix, and the anonymized persona uses the final anonymized pair. Each persona answers each probe in five rounds at temperature 0.7, and User A's probes use three rounds at temperature 0. All replies of both personas to a probe are then read side by side and compared on how each persona reacts to the violations: tone, sentence shape, and acts such as correcting a fact, restating a constraint, or handing the task back. A probe passes this reader check when both personas respond as the same kind of person, even when the wording differs.

Table 6 reports the results for each user. All 14 probes pass the reader check.

Shortcut check. We next test whether anonymization removes the shortcut of guessing a user's preferences from prior knowledge of the source character. A PA backbone predicts the user's full 2×14 preference matrix from a prompt that lists the 14 attributes with their settings and asks for one setting per attribute in each context. The prompts differ only in what they say about the user. In the source-identity condition, the prompt names the character and the show (The Big Bang Theory or the U.S. The Office) and adds the original profile. The anonymized-profile condition gives only the anonymized profile, and the no-information condition gives nothing about the user. A profile covers the occupation, living arrangement, and relationships, and it never includes personality or speech style. Accuracy is the fraction of the 28 cells whose predicted setting matches the user's final preference matrix. Chance accuracy is the mean over the 28 cells of one over the number of settings of that attribute, which is .452. We run the four PA backbones of the main experiments, GPT-5.6-Sol, Qwen3.8-27B, HY4-Preview, and DeepSeek-V4-Flash, on all four users with three repeats at temperature 0.7 and reasoning turned off. The no-information prompt is the same for every user, so each backbone's three predictions in this condition are scored against all four matrices.

Figure 5 shows the accuracy for each user and overall. With the source identity, accuracy reaches .532, which is .080 above chance. With the anonymized profile, accuracy falls to .464, which is close to the no-information condition at .462 and to chance at .452. The source identity gives the highest accuracy for every user. The anonymized profile is above the no-information condition for Users B and C and below it for Users A and D. Knowledge of the character therefore helps the prediction, and anonymization removes this help.

B.5 Worked Extraction Example

We trace one preference cell of one source character from the transcript scenes to the final matrix. The cell is the work Verbosity setting of User C, which is Terse, and it shows how the second pass decides between the character's own speech and the character's responses to others (Appendix B.3).

Scenes. The first pass tagged four work scenes of the character for Verbosity. Two of them show the character's own speech. When a visiting saleswoman asks for a coffee, the character turns the

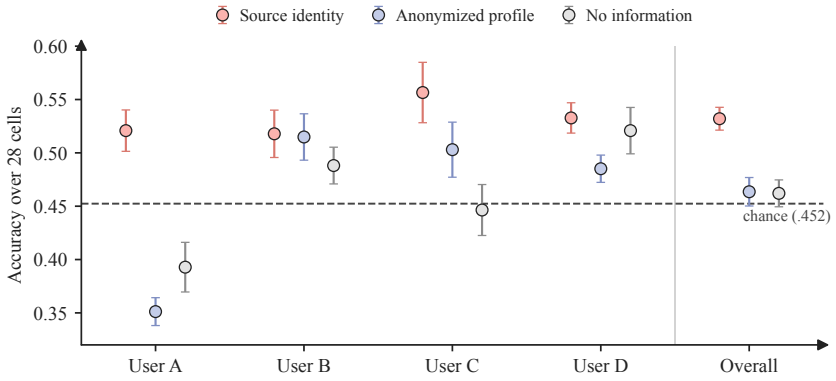


Figure 5: Shortcut check. Accuracy of predicting the user's 28 preference cells when the prompt gives the source identity (character, show, and original profile), the anonymized profile, or no information about the user. Points are means over four PA backbones and three repeats, that is, 12 runs for each user and 48 runs overall, and error bars are standard errors of the mean. The dashed line marks chance accuracy, the mean over the 28 cells of one over the number of settings of that attribute (.452).

request into a long riff about a coffee chain and about being a nonconformist before asking how the coffee should be made (season 1, episode 6). When a supervisor moves the character's meeting to the warehouse, the character turns the move into a grand speech about the hidden world below the office (season 2, episode 15). The other two scenes show responses to others. In a staff meeting about a holiday, the character asks for its origins, and a subordinate answers with what the holiday celebrates, the battle behind it, and what it symbolizes. The character cuts the answer off and says that this is not an epic fantasy saga (season 3, episode 6). In the other scene, a subordinate reports in one sentence that the whole office has been interviewed about an absent colleague, and the character tells the subordinate to go to the colleague's house and ends the exchange with a two-word order (season 2, episode 13).

Second pass. The second pass set the cell to Terse with high confidence. Its summary notes that the character gives long, theatrical monologues when running a meeting but cuts long explanations short when receiving them. The two kinds of evidence point to different settings, and the responses to others decide.

Review. When we checked the responses cited for the high-confidence settings of Users C and D, the second response did not hold. The update it answers is already short, and the order may come from urgency to get the task done, so it does not show how the character wants to be spoken to. With one response left, we lowered the cell to medium confidence and reviewed it with the others. The competing setting was Detailed, which the character's own long speeches point to. The remaining response decides against it. The character asked for the origins and disputes none of the answer, so what the character stops is the elaborate telling. This response reacts to style, and we kept Terse.

Final fields. The final matrix of User C records the cell as follows.

```
work:
  verbosity:
    setting: Terse
    source: extractor_accepted
    confidence: Medium
```

The source label records that we kept the extracted setting, and the medium confidence reflects the single response to others. The personal Verbosity setting of User C is Detailed, based on the character's long anecdotes and speeches outside work, so User C wants short answers at work only. The identity keeps the long delivery in its speech style, and the skill for Terse (Appendix D.2) lets the user's own turns run longer than the answers the user accepts. The simulator therefore still speaks at length as User C while it expects short answers from the PA at work.

C World and Session Construction

C.1 World Design, Timeline, and Fixtures

Each user's world covers four months of their work and personal life. Its storyline is written for the user, with recurring people and places that the user regularly meets and visits, and it has six scenario types for each context, from which the session tasks are seeded. A coding agent wrote the Timeline and World Design of each user in the same structure, after we chose the five changing cells of the user and their new settings. Scripts check that each Timeline lists the 94 accumulation sessions and five change events in time order with the events at fixed positions, and that the changing cells in each World Design match the Timeline and the preference matrix. The three accumulation sessions of a stable cell have different tasks, and accumulation sessions average 5.5 user–PA exchanges and 14.6 tool calls. For the fixtures, a coding agent wrote the records after we approved its world canon, which fixes every name, date, and number, and a script checks that each record a decision point needs is visible on the session date.

C.2 Tools and World State

The PA reads and changes the world state W_t only through tool calls. MemPA provides 83 tools across 26 domains, and Table 7 lists them by domain. The 51 read tools search and read records, and the 32 write tools create, update, or delete them. Eight write tools change the user's own records, such as documents and calendar events. The other 24 act toward other people or services, for example by sending a message or paying a bill. The tools run in a separate server that the PA reaches through the Model Context Protocol (MCP) (Anthropic, 2024), and the PA sees each tool under the prefix `mcp_state_`, as in `mcp_state_calendar_event_create`. Every PA receives all 83 tools in every session. A few domains hold records only in the worlds of the users whose storylines need them, such as exhibitions for User A.

Each user's world starts from a fixture tree, which is a directory of plain files built during world construction (Appendix C.1). These files hold the user's emails, messages, calendar, contacts, and documents, together with the records of the other domains, such as bills, orders, and open booking slots. When a run starts, the whole tree is copied into the run, and this copy is the persistent world of that run. A run pairs one user with one backbone and one memory condition. Each run therefore changes only its own world, and the worlds of different runs diverge as their PAs take different actions.

Each accumulation session, change event, and probe runs as one step on this world. Before a step starts, the persistent world is copied into a working copy, and the tool server of that step reads and writes only this copy. During the step, the files of the working copy are the world state W_t. The write calls that run in turn t change these files, and this change is the transition $\mathcal{T}$ from W_t to W_{t+1}. A write that the IX wrapper holds for the user's approval does not run, so it changes nothing until it is released (Appendix E).

A step ends in one of two ways. After an accumulation session or a change event, the working copy is committed and replaces the persistent world, so its changes carry into all later sessions. A change event replays its fixed tool calls through the same tools, so its actions enter the world in the same way as the PA's own calls (Appendix C.4). A probe starts from the world saved at its checkpoint, and its working copy is discarded when it ends. Nothing a probe does therefore reaches the persistent world or another probe (Section 3.4). A step that fails is also discarded, so the persistent world holds only the changes of completed steps.

Every tool is a fixed function of the world files and its arguments, and no answers are scripted. A read tool loads the relevant files and returns the matching records. A write tool changes the files, and later reads return what it wrote. For example, `calendar_event_create` adds an event to the calendar file, and `calendar_event_search` returns this event later in the same session and in every later session of the run. Each tool call is logged, and each call that changes the world also records the change. These logs are kept with the output of the step, so every change to W_t can be traced to the call that made it. An accumulation session with its tool calls and state changes is given in Appendix C.6.

The tools also apply two visibility rules. The first rule concerns time. Records that appear during the timeline carry the time they enter the world, such as the date an email arrives, and the tools hide

Table 7: World tools by domain. n is the number of tools in each domain. Read tools are set in regular type and write tools in bold. Write tools marked † act toward other people or services, and the other write tools change the user's own records. The PA receives all 83 tools in every session, each under the prefix `mcp_state_`.

Domain	n	Tools
Email	6	`email_message_search`, `email_message_read`, `email_draft_read`, `email_sent_read`, **`email_draft_create`**, **`email_message_send`**†
Messages	3	`message_search`, `message_read`, **`message_send`**†
Calendar	4	`calendar_event_search`, `calendar_event_read`, **`calendar_event_create`**, **`calendar_event_update`**
Documents	6	`copier_entitlement_read`, `document_read`, **`corporate_reporting_submit`**†, **`document_create`**, **`document_update`**, **`document_delete`**
Contacts	4	`contact_search`, `contact_read`, **`contact_create`**, **`contact_update`**
Phone	2	`phone_call_read`, **`phone_call_dial`**†
Maps and places	3	`club_closure_calendar_read`, `map_route_search`, `place_search`
Weather	1	`weather_forecast_read`
Shopping	6	`shopping_product_search`, `shopping_order_read`, `shopping_return_case_read`, **`shopping_order_create`**†, **`shopping_order_update`**†, **`shopping_return_case_update`**†
Shipping	3	`delivery_address_read`, `package_delivery_proof_read`, `package_tracking_read`
Finance	5	`utility_bill_read`, `payment_method_read`, `bill_payment_read`, **`utility_account_holder_update`**†, **`bill_payment_create`**†
Appointments	4	`appointment_search`, `appointment_read`, **`appointment_create`**†, **`appointment_update`**†
Restaurants	4	`restaurant_reservation_search`, `restaurant_reservation_read`, **`restaurant_reservation_create`**†, **`restaurant_reservation_update`**†
Room booking	4	`room_booking_search`, `room_booking_read`, **`room_booking_create`**†, **`room_booking_update`**†
Hotels	4	`hotel_search`, `hotel_booking_read`, **`hotel_book`**†, **`hotel_booking_update`**†
Reimbursement	2	`reimbursement_submission_read`, **`reimbursement_submission_create`**†
Conference registration	2	`conference_registration_read`, **`conference_registration_create`**†
Grade submission	2	`grade_submission_read`, **`grade_submission_create`**†
Flights	4	`flight_search`, `flight_read`, `flight_booking_read`, **`flight_book`**†
Scholarly papers	2	`scholarly_paper_search`, `scholarly_paper_read`
Recipes	1	`recipe_search`
Comic store	1	`comic_store_inventory_read`
Cinema	3	`cinema_showtime_search`, `cinema_ticket_read`, **`cinema_ticket_purchase`**†
Exhibitions	3	`exhibition_search`, `exhibition_ticket_read`, **`exhibition_ticket_purchase`**†
Media transcoding	3	`media_transcode_service_search`, `media_transcode_job_read`, **`media_transcode_job_create`**†
Household	1	`household_inventory_search`
Total	83	51 read tools and 32 write tools, of which 24 are marked †

every record that enters after the date and time of the current session. One fixture tree can therefore hold a user's whole timeline, and a session early in the storyline does not see an email that arrives months later. The rule depends only on when a record enters the world, so an appointment that was booked last week for next month stays visible. Documents can mark their sections with dates in the same way, and a session then reads only the sections written by its date. The second rule concerns paths. The PA can reach only files inside its working copy, and the construction notes stored in the fixture tree and the step logs stay hidden from it.

Each tool checks its arguments before it reads or writes. The parameter descriptions that the PA sees state the format of each value, such as `YYYY-MM-DD` for dates and 24-hour `HH:MM` for times, and

the tool rejects any value in another format. A value that names nothing in the world is rejected in the same way, such as an unknown city or a booking slot that does not exist. A rejected call writes nothing. Its error message returns to the PA as the tool result and says which value failed and why. For example, a calendar call with the date `2026-11-6` returns the error that this value is not `YYYY-MM-DD`. The PA can then correct the value and call the tool again.

C.3 From Outline to Accumulation Sessions

Each accumulation session, probe, and change event is written as a script made of beats. A beat is one step of the session from the user's side. It gives the simulator the user's goal at that point, what to say, the condition under which the beat starts, and the world records it refers to. The beat in which the tested attribute becomes active is the decision point. This beat also describes how the user reacts under each setting of that attribute and what the user does next when the PA's behavior falls short of the goal.

Reaching the decision point. Four mechanisms keep a live session on its path to the decision point, and they apply alike to accumulation sessions and probes. Change events follow fixed transcripts and need none of them (Appendix C.4). First, each task fact the user holds is assigned to the beat at which the user reveals it. Before that beat, the fact is withheld from the simulator itself, so it cannot be given early, and a simulator asked for it early says it will give it at that step. Second, construction checks reject a draft in which an earlier beat already poses the kind of request that the decision point tests, or in which the opening request already contains the missing information. These checks cover all 14 attributes. Third, for five attributes (Reasoning Visibility, Process Visibility, Autonomy Level, Information Elicitation, and Topic Management), the task ends in a write whose content the user dictates at the decision point, and the state server refuses this write until the user has provided the content. The PA can then proceed only by involving the user. The other nine attributes are tested in a short exchange of the decision point and one reaction, which needs no such lock. Fourth, the interaction tool of a must be called at the decision point. A PA that omits the call is reminded once, and the call is then forced. If it still fails, the step ends with no scorable reply and no state change, and it may be retried.

The prompts that generate these scripts are summarized in Appendix G.2.

C.4 Controlled Preference-Change Events

The storyline template spaces six event slots along the timeline, three work slots followed by three personal slots. Each user uses the three work slots and two of the personal slots, which gives the five change events of Section 3.2. Each slot also fixes where the three accumulation sessions before its event and the two after it take place.

For each event, we choose a cell whose change fits the user's storyline. The old setting is the cell's setting in the user's preference matrix, so besides the cell we choose only the new setting and the slot. Proactive Outreach, whose decisions are triggered by events in the world, has no change events. The 20 events cover nine attributes, and the changes run in both directions. For example, Solution Breadth changes from Low to High for Users A and D and from High to Low for User B.

Each event follows one causal order. The PA first carries out the old setting in full, and this behavior has a consequence that the user meets in the conversation. The user then sees the behavior as the cause and, in the last message of the event, states the new rule in their own words. The consequence follows from the tested behavior only, so the PA still completes the task well and the event gives no reason to change any other cell. The change-event generator writes a draft in this order, and the transcript writer turns the draft into a fixed transcript (Appendix G.2).

The event is replayed from its fixed transcript into the world, the PA's history, and its memory through the same interface as a live session (Section 3.3). The evaluated PA generates nothing during the event, and the simulator does not run. For the simulator, each event comes with a short first-person note of what happened and what the user now wants, and this note is added to the simulator's memory. From the event on, the target cell takes the new setting and every other cell keeps its setting. The pre-event probe of the cell runs one hour before the event, and PET pairs it with the final probe of the same cell (Section 3.4). A worked example is given in Appendix C.6.

Table 8: The five conditions of setting viability. A setting in $\mathcal{P}_a$ is viable in a session when the PA follows it and all five conditions hold. *Rules* are the rule-based validators, and *Review* is the human-verified LLM-based semantic review, which reads every draft against all five conditions.

Condition	Meaning	How it is checked
1. Feasible	The PA can carry out the setting with the records and tools in W_t.	*Rules:* the needed records are bound, exist, and are visible in time; the tool steps run on a scratch copy of the world. *Review:* the records hold what each setting needs.
2. Reasonable	An assistant that does not know the user's preference could choose the setting, and the choice is neither careless nor harmful.	*Review:* under each setting, the reply is a complete and legitimate answer to the request.
3. Not given away	The user's messages do not name the setting or ask for it.	*Review:* no user message before the decision point asks for a setting or cues one. *Rules:* e.g., a Solution Breadth probe request states no number of options.
4. Observable	The settings lead to interactions that differ naturally in the PA's reply or executed tool calls, and each setting's required behavior is done in full.	*Rules:* the task has the structure the attribute needs, e.g., two dependent tool calls for Process Visibility. *Review:* the replies under different settings differ visibly and naturally.
5. Answerable	In the next beat, the user has a way to respond to each setting.	*Rules:* the next beat carries a user reaction, and a probe states the expected reply under every setting. *Review:* each setting has its own reaction path.

C.5 Construction Checks and Acceptance

Each accumulation session, probe, and change event tests one attribute a at a decision point. We call a setting in $\mathcal{P}_a$ *viable* in a session when the PA follows that setting and all five conditions in Table 8 hold. These conditions turn the decision-point requirements of Section 3.1 into checks on each draft. Conditions 1 and 4 require that every setting is feasible with the records and tools in W_t and leads to an observably different interaction. Conditions 2 and 3 require that no setting is wrong for a PA that does not know the user, and that the user does not name the setting. Condition 5 adds that the user has a way to respond under every setting, since the user's reactions are how a preference reaches the PA in accumulation sessions.

In condition 4, an observable difference is one that appears in the text the judge reads, which is the PA's public reply and the task tool calls that actually ran (Section 3.4). The difference must also come naturally from the task, so a draft fails when its settings differ only because of a detail added to force them apart. Each setting's required behavior must be done in full, but a setting need not finish the original task. Under Capability Boundary, *Suggest alternatives* passes once the PA tries a further route after the block, even if that route also fails. *Find and hand off* passes when the PA makes a valid hand-off. A valid hand-off states what was not done, why it is blocked, and what the user needs to decide or do next. It never claims success. It may include the reads needed to explain the block, but it makes no further attempt to recover.

The viability check gives each draft one of two outcomes. A draft is *live* when every setting in $\mathcal{P}_a$ is viable, so each setting can be carried out reasonably, the settings differ naturally and observably, and the user can respond to each of them. A draft is *closed* when any condition fails for any setting, and the reviewer records the specific reason. For example, two settings may lead to the same interaction, the user may have no way to respond to one setting, or the difference between settings may be contrived.

Two kinds of checkers apply to accumulation sessions, probes, and change events. Rule-based validators run first and check what a program can decide from the draft and the world. They check that each draft binds the world records its attribute needs, for example one candidate record for each option under Solution Breadth, and that these records exist and are visible by the session date. They also run the draft's tool steps in order on a scratch copy of the world, check the order of its beats, and

require the beat after the decision point to carry a user reaction. Probes must in addition describe the expected reply under every setting in $\mathcal{P}_a$, and each tool call in a change event is replayed on a scratch copy of the world. Separately, tests of the tool server check that each tool matches its catalog entry and reads and writes the world state correctly.

Rules cannot decide whether a setting is a reasonable choice or whether a difference comes naturally. An LLM-based semantic review therefore reads every draft against all five conditions. Its golden criteria come from one written specification of viability, which states the five conditions and, for each attribute, what each setting needs from the world. The review rubrics for accumulation sessions and probes point to this specification. The generator of accumulation sessions receives the same specification, so each draft is written against the conditions that later check it. The rubrics also make the conditions concrete. The accumulation rubric, for example, asks whether each setting has its own reaction path in the next beat. The review is run by LLM agents (Codex and Claude Opus). A second model is used only to catch what the first missed, and it applies the same criteria. We then read the drafts and the verdicts ourselves.

Change events need one adjustment. The PA's turns in a change event are fixed and follow the old setting $p^{(t_k^-)}_{u,c_k,a_k}$. We check the old setting under the same conditions as in an accumulation session, and the review confirms that the PA carries it out in full before the consequence appears. The new setting is not played out in the event, so for it we check only that the user has a reaction to it and that the task and its starting facts would stay the same. One timing rule applies only to change events. When a write needs a value that only the user can give, the user gives it before the beat in which the PA acts on the old setting. Otherwise the write would be blocked, and the block could be mistaken for the behavior the event exists to show.

A draft is accepted only when it passes the rule-based validators and the semantic review finds it live. A draft that fails a rule goes back to generation before any review. We revise a closed draft by regenerating or editing it, with the recorded reason as the guide, and then run the validators and the review again. This loop ends only when the draft is live. Each accepted draft is fingerprinted, so a later edit voids its acceptance, and only accepted drafts enter a run.

C.6 Accumulation Session and Change Event Examples

In a work accumulation session of User B on Autonomy Level with target *Self-directed*, the PA selects *Autonomous* and sends a text to Carla with `message_send`, which adds message `msg_0060` to the world state. User B replies that "anything going out to someone else needs to be run past me first." In User B's personal change event on Topic Management, the PA handles three errands in one reply under the old setting *Organize*, a cheap red-eye flight stays unbooked, and User B states the new setting *One-at-a-time*: "work on just the first item and keep the rest waiting."

D User Simulator and Preference Exposure

D.1 Turn Execution, Progression, and Termination

In each turn, the simulator writes its message and private blocks, and only the message reaches the PA after the message checks pass. The PA answers with its memory evidence and tools, the wrapper renders the answer, and the simulator reads the delivered reply with a private record of the actions that ran and the items held for its answer (Appendix E.4). The session moves to the next beat after the simulator's beat decision marks the beat's goal as met, and it ends after the last beat, which for the nine attributes tested in a short exchange is the reaction to the decision point.

D.2 Preference Exposure through Beats and Active Skills

The simulator shows a preference through a *preference skill*. Each of the 32 settings has one skill, written for the setting and shared by all four users. A skill describes the behavior that the user wants from the PA and the signals in a reply that go against it, and it has two sides. The production side shapes the user's own turns. The reception side says how the user takes a reply of the PA: which reply follows the setting, and which move the user makes when a reply goes against it. The wording and intensity of each turn come from the user's identity, so one skill sounds different for User A and for

User C (Appendix B.5). Under the skill for Terse, for example, an answer in the first sentence with at most one supporting move is right, and a reply that stacks background and caveats behind the answer leads the user to pick out the action point and leave the additions unaddressed.

Skills are tied to the beats of the session script (Appendix C.3). Each beat declares its active attributes, and the decision point declares the tested one. On a declaring beat, the simulator prompt holds the full skill of each active attribute at the user's current setting in the session's context. On the next beat, where the user reacts to the PA's reply, the prompt keeps only the reception side, and after that beat the skill is dropped. The PA selects its setting once at the decision point, so the user reacts to the delivered interaction once and then continues the task in the manner the PA chose. Probes load the same skills, but only the production side applies, and the user gives no feedback on the manner of a reply (Appendix G.3).

The reception side leaves no room for silent acceptance. On the first turn after a reply of the PA shows the behavior that a skill is about, the user takes a position on that behavior in the same message as the task. When the reply follows the setting, the user says what they take from it, and a general "fine" does not count. When the reply goes against the setting, the user makes the move that the skill names, such as pushing back, correcting, asking again pointedly, or answering shorter. Before the behavior appears, the user plays the beat as written. The position takes one or two clauses and never names the preference (Appendix D.4). A correct selection therefore leaves an explicit acceptance in the conversation and a wrong one leaves a correction, so both give the memory a signal to store.

Every accumulation session reaches this reaction. Construction checks require a reaction to every setting and require the beat after the decision point to carry it (Appendix C.5). At run time, the PA must call the interaction tool of the tested attribute at the decision point (Appendix C.3). A beat that declares skills is never skipped, and neither is the beat after it, and a beat that runs out of turns moves the session ahead toward the next beat that declares skills. Two annotators confirmed on a sample of accumulation sessions that the reactions convey the preference (Appendix F.9).

D.3 Cross-Session Continuity and Information Boundaries

Three stores carry over from one session to the next. The world keeps every write that the PA's tools carried out, and its time rule hides records that enter after the current session (Appendix C.2). The PA's memory receives each turn as the PA saw it through the common adapter (Section 3.3). The simulator keeps a memory of its own. After each session, it writes a first-person entry of two to four sentences, dated in the story's time, on what it asked for, how the PA handled it, what it decided, and what is still open, and it updates a short record of its lasting impressions of the PA. The entry leaves out the subject matter, file paths, and exact numbers and dates, so the simulator remembers how the PA treated it, and the task content of each session comes from its script. A change event adds a first-person note of the change to this memory (Appendix C.4).

Table 9 lists what each component can see. The PA meets the user's preferences only in the user's messages and in what its memory returns. The simulator reacts to the PA's public reply and the actions that actually ran, and it never sees the PA's selection or the content that the wrapper withholds. The memory never receives the content that the wrapper withholds, and both realization scores are given without the selection. Probes branch from a pinned checkpoint and write nothing back to the world, the PA's memory, or the simulator's memory (Appendix F.2).

D.4 Preference Leakage Control and Audit

The PA should learn a user's preferences only from how the user reacts to its behavior. The simulator would leak a preference if it gave the preference away by another route, for example by naming the setting. We prevent leakage at run time with prompt rules, a split between private and visible output, and rule-based checks. After the experiments, we audited the simulator messages of the formal runs for any leakage that got through.

Prompt rules. The simulator prompt tells the simulator to show each active preference through behavior and never to declare it. The prompt lists each active attribute with its setting, so the simulator knows what to act out, and it adds that a reaction must never use the name of the preference or the attribute. The global rules also forbid statements such as "I prefer X" and any mention of the

Table 9: Information available to each component during a session. ✓: available; ◐: partly available, as noted below the table; ✗: not available.

Component	Profile and settings	Session script	Simulator memory	PA reply and executed actions	PA selection and withheld content	Memory evidence
User simulator, accumulation	✓	✓	✓	✓	✗	✗
User simulator, probe	✓	✓	✗	✓	✗	✗
PA	✗[a]	✗	✗	✓	✓	✓
Memory system	✗	✗	✗	✓	◐[b]	–
BRS judge	◐[c]	◐[d]	✗	✓	✗	✗
PRS rating	◐[e]	✓	✗	✓	✗	✗

[a] Under Oracle, the PA also receives the target setting of the probed cell. [b] The PA's call to the interaction tool, without the content that the wrapper withholds. [c] The target setting of the probed cell with its definition. [d] The probe's task and the world state at its start. [e] The persona card and the acting sheet of the probed setting.

benchmark, the script, the beat, the taxonomy, or the simulator. A preference therefore shows only in what the user asks for, accepts, rejects, or corrects. The prompt further ties every reaction to the PA's actual last move. The simulator may not react to a behavior that has not appeared in the PA's replies, and it leaves the shape of each answer to the PA. Its requirements thus come out as reactions over the course of the conversation and never as a checklist at the start. Probe sessions use a separate prompt. This prompt tells the simulator that it has never spoken to the assistant before, and it forbids phrases that point back to an earlier session, such as "last time", "as usual", or "you remembered". Both prompts are summarized in Appendix G.3.

Private blocks. Each simulator turn has one visible block and two or three private blocks. The visible `<message>` block holds what the user says to the PA. The private `<factual_check>` block records factual errors of the PA, and the `<turn_assessment>` block decides whether the session moves to the next beat. On turns with held actions, the `<consent_decisions>` block records whether the user allows each one. Only the text inside `<message>` is passed to the PA.

Rule-based checks. Before a message reaches the PA, rule-based checks look for text that comes from the simulator's instructions or private notes. They look for private-block tags, the consent block, the benchmark's own vocabulary such as "target setting" or "probe polarity", and beat identifiers from the session script. A failing message goes back to the simulator with a note that names the problem, and the simulator writes the turn again. A message that still fails after three retries is never delivered, and the session is marked invalid. In the audited runs, the checks rejected 43 generated messages, and none of the 21,093 delivered messages fails them.

Post-hoc audit. After the experiments, we audited the simulator messages of the formal runs. The audit covers 34 accumulation runs with 3,193 sessions and 111 probe runs with 3,377 probe sessions, which contain 31,703 and 17,051 simulator sentences. A rule filter first flags candidate sentences. It matches the names and labels of the 14 attributes, setting names that appear with a preference phrase such as "I'd rather" or "next time", phrases of a standing preference or of a remark on the current reply, and, in probes, references to an earlier session such as "last time". To see what the rules miss, we also read a fixed random sample of unflagged sentences, and we swept all sentences for the frequent request to check with the user before sending. Claude Opus 5.5, running as a Claude Code agent, then read every flagged, sampled, and swept sentence together with the start of the PA's preceding reply. It labeled each sentence as a remark on the current reply, a general request that follows from a reaction, or neither, and it marked any probe sentence that refers to an earlier session. We then reviewed about 300 of these sentences by hand, including every sentence that the reader marked as possible leakage. The findings of this review refined the rule filter and the run-time checks, and the formal runs use the refined versions.

Criteria. We count a sentence as feedback when it reacts to something the PA has just done in the same session. Feedback is how the benchmark exposes a preference, so it is not leakage. This holds for a remark on the current reply. It also holds when the user turns the reaction into a general request, for example by asking the PA to check before it sends anything, right after the PA has sent a message on its own. Leakage takes one of two forms. In the first, a probe sentence refers to an earlier session.

Table 10: Post-hoc leakage audit of simulator sentences. Each row gives the sentences read from one source in accumulation or probe sessions. A remark on the reply comments on the PA's current reply, and a rule-like reaction turns a reaction to the PA's last action into a general request. Other covers task content and chance rule matches. These three labels sum to the judged count. Leaks counts probe sentences that refer to an earlier session and sentences that state a preference or name a setting with no trigger.

Source	Phase	Judged	Remark on reply	Rule-like reaction	Other	Leaks
Rule hits	Accumulation	698	621	5	72	0
Rule hits	Probe	375	261	3	111	0
Random sample	Accumulation	300	39	0	261	0
Random sample	Probe	150	8	1	141	0
Found in context	Accumulation	5	0	5	0	0
Found in context	Probe	2	0	2	0	0
Phrase sweep	Accumulation	238	89	2	147	0
Phrase sweep	Probe	71	14	3	54	0
Total		1,839	1,032	21	786	0

In the second, the simulator announces a preference, or names a setting or its definition, with no trigger in the PA's behavior.

Results. The audit found no leakage. The counts are given in Table 10. The reader judged 1,839 sentences, and 598 of them come from probes. No probe sentence refers to an earlier session, and all 19 sentences flagged by the earlier-session rule are false matches. No sentence announces a preference or names a setting with no trigger. Of the judged sentences, 1,032 remark on the PA's current reply, and 21 turn a reaction into a general request. Each of these 21 sentences reacts to an action the PA had just taken. All 21 come from Users B and D, and 17 of them ask the PA to check with the user before a message goes out to someone else. The other 786 sentences concern the task or match a rule by chance, for example an instruction on the tone of a text that the user sends to another person.

E PA Interaction and IX Wrapper

E.1 Purpose and Position in the Execution Flow

The IX wrapper sits between the assistant's agent loop and the public interaction (Figure 6). Its input is the assistant's selected settings together with assistant-authored response material and proposed task actions. Its output is one user-visible reply plus the results of actions that were allowed to execute or the status of items held for a later turn. The wrapper has no access to the target setting, so it standardizes how a choice is realized without correcting the assistant's choice.

E.2 Setting Selection and Subsequent Behavior

An IX selection tool becomes available when its attribute first activates in a session. The assistant calls it once, and the selected setting remains active for the rest of that session. The same tool is used in accumulation sessions and probes. Before the call, the wrapper passes the assistant's output through unchanged. Later turns do not reselect the setting; instead, they provide turn-level payloads such as rationale text, topic batches, clarification questions, outreach content, or proposed actions. This separation keeps the evaluated choice explicit while allowing the public interaction to respond to changing task state.

E.3 Wrapper Mechanisms and Activation

The wrapper uses four processing mechanisms. Pass-through retains the assistant's wording for Tone & Formality, Verbosity, Emotional Engagement, Guidance Level, Solution Breadth, and Task

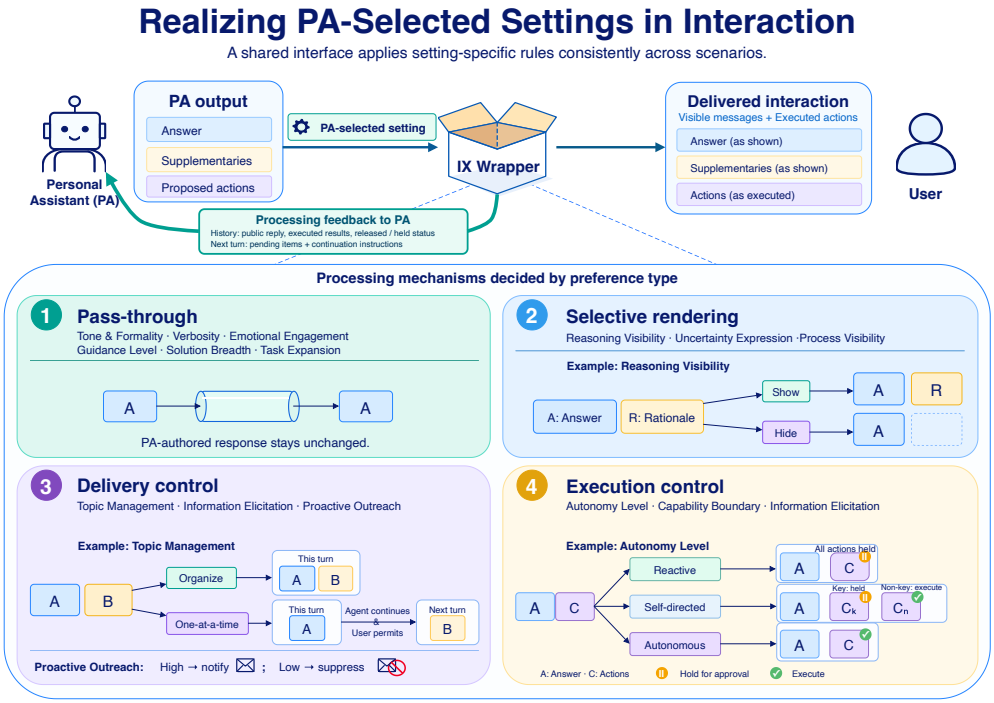


Figure 6: IX wrapper boundary and processing mechanisms. The personal assistant selects an interaction setting and supplies turn-level content or action payloads. The wrapper applies pass-through, selective rendering, delivery control, or execution control without access to the ground-truth preference. It emits one public reply together with executed results or held-item status; held content and processing traces stay in the audit record and never enter assistant memory.

Expansion. Selective rendering controls assistant-authored rationales, uncertainty statements, and process notices for Reasoning Visibility, Uncertainty Expression, and Process Visibility; provider-private reasoning is excluded from all public fields. Delivery control schedules topic batches, clarification questions, and proactive outreach for Topic Management, Information Elicitation, and Proactive Outreach. Execution control gates task actions for Autonomy Level, Capability Boundary, and Information Elicitation.

For Autonomy Level, *Reactive* holds substantive actions for approval, *Self-directed* executes drafts and routine record changes but holds key record changes and every outward action, and *Autonomous* executes actions within the delegated task. Held material retains an item identifier and continuation state so that later consent or information releases the intended item rather than an unrelated action.

E.4 PA Output, Public Interaction, and Memory Records

The simulator receives the rendered public reply and the consequences of actions that actually executed. The assistant receives runtime feedback describing executed and held items so that it can continue coherently. The memory receives the user message, the PA's tool calls with their results, its call to the interaction tool, and the delivered reply. Held content, policy fields, and rendering traces are kept in an audit record for scoring and debugging and are never written to memory. The behavior judge compares the public text and executed-task trajectory with the target setting; it does not receive the assistant's IX selection or internal wrapper policy fields.

E.5 Worked Wrapper Example

In a personal accumulation session, User D asks the PA to put Thursday art classes on the calendar and then email an enrollment request to the studio. The PA selects the target setting *Self-directed*,

Table 11: Memory conditions under the common MemPA adapter. Each condition receives the same record of each turn: the user message, the PA's tool calls with their results, its call to the interaction tool, and the delivered reply.

Condition	Write	Retrieval
File memory	LLM consolidation into `MEMORY.md` and `HISTORY.md`	Consolidated Markdown supplied to the assistant
Simple RAG	Raw messages stored through Mem0 with `infer=False`	Vector retrieval over raw messages
Mem0	Memory extraction with `infer=True`	Native Mem0 search over extracted memories
MemOS	Native MemCube add operation	Native MemCube search operation

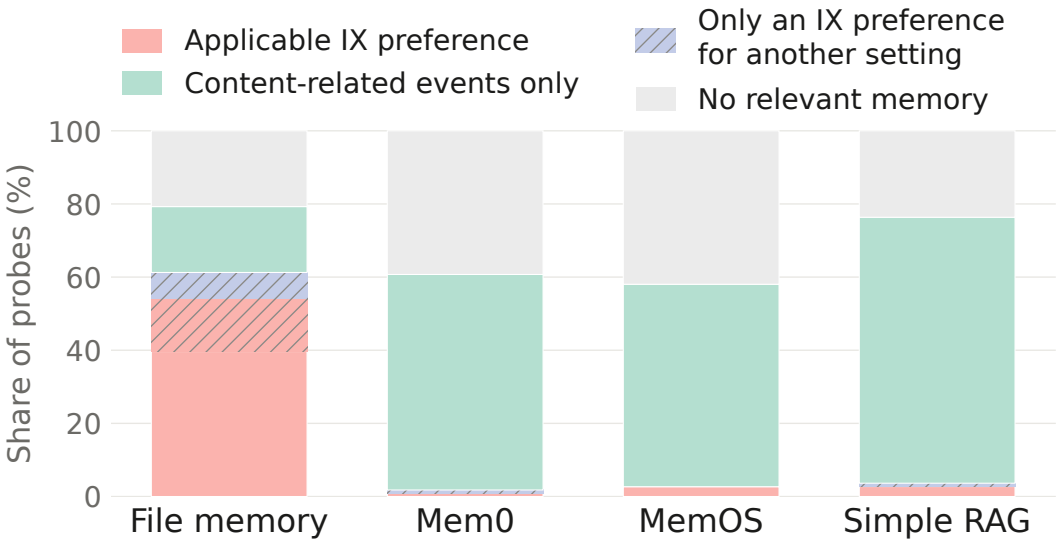


Figure 7: What the PA sees before its first selection, in 112 sampled final probes per memory condition.

so the wrapper lets the calendar write run, holds the `email_message_send` call, and ends the reply with the PA's question "Send Ines Varela the enrollment request [...]?" User D answers "Yes, go ahead and send her the enrollment request," the PA releases the held item, and the email goes out as `sent_0001`.

F Evaluation Protocol and Validation

F.1 Memory Conditions and Comparison Controls

Table 11 lists the four conditions and their backend-specific operations. The common adapter writes each turn to memory as the PA saw it: the user message, the PA's tool calls with their results, its call to the interaction tool, and the delivered reply. File memory consolidates this history into `MEMORY.md` and `HISTORY.md`; Simple RAG stores raw messages through Mem0 with `infer=False`; Mem0 uses `infer=True` extraction; and MemOS uses native MemCube add and search. For each model, all conditions share the same users, session schedule, wrapper, tools, checkpoint timepoints, and probe protocol. Their accumulated histories and memory states remain condition-specific. Assistant and simulator memories use separate stores.

Figure 7 shows what each memory condition brings into the PA's input before its first selection, in 112 sampled final probes per condition (Finding 2).

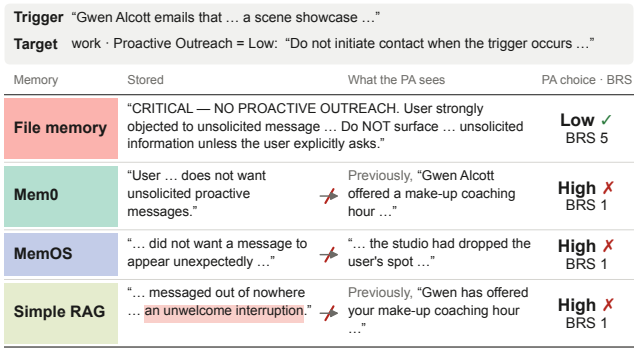


Figure 8: Stored memory and what the PA sees before one outreach decision (User B, work, target Low, GPT-5.6-Sol). All four stores hold the user's objection, but only File memory shows it to the PA.

Figure 8 shows, for one final probe of User B, what each store holds and what it brings into the PA's input before its selection (Finding 3).

F.2 Probe Construction and Execution Isolation

Each probe repeat begins from a frozen checkpoint and runs on an isolated branch. Final probes cover the 28 attribute–context cells for each user. A pre-event probe is paired with the final probe for the same changed cell and the same repeat index. Probe interactions cannot write back to the accumulation trajectory, so one repeat cannot alter the memory or world state used by another.

F.3 Scoring Inputs, Rules, and Aggregation

PSA compares the assistant's selected setting with the frozen ground truth by exact match. A response that is present but lacks a valid IX selection is scored incorrect. Unresolved missing probe or scoring artifacts are marked incomplete and are not converted to zero. BRS is the independent 1–5 LLM Judge score computed by comparing the target setting with the public text and executed-task trajectory. The judge is blind to the assistant's IX selection and internal wrapper policy fields. For PRS, the user simulator keeps its persona card and model but not its accumulated memory, and it gives a short reason with each rating. It sees the public exchange and its own acting sheet for the probed setting, but never the PA's selection or the judge material. PRS is aggregated as BRS over the same 28 final probes per repeat.

For a probe with a documented model-capability failure, we assign PSA $= 0$ and BRS $= 1$ without reducing the denominator; any CS or PET pair containing the failed endpoint does not pass. These protocol assignments represent failure handling rather than observed LLM Judge outputs. A probe is floored only when the PA's own output breaks the interaction-tool contract. When the PA leaves out a required interaction-tool call, the harness reminds it within the turn and, if the probe still fails, runs the probe again, at least three runs in all. Of the 70 floored probes (67 final, 3 pre), 64 failed all of their runs, three runs for 58 and five or six for 6. In 60 of these, the PA left out a required interaction-tool call, usually the turn-level payload after its selection, and in 4 it held an action without asking the user to confirm it. The other 6 failed in ways that are not rerun: in 4, the PA wrote interaction-tool text into a task tool's arguments, which would show its selection to the blind judge, and in 2 it made no send-or-suppress decision on a proactive event. Table 12 gives the final floors by condition and backbone. Under Oracle, 5 of the 13 final probes that score 0 are floored, and in all 5 the PA had selected the target setting before the failure; the other 8 are wrong selections.

One Oracle final probe (User B, GPT-5.6-Sol, repeat 1) could not be faithfully rebuilt for self-report and is excluded from PRS, so that repeat's PRS averages 27 of its 28 final probes (its BRS keeps all 28).

CS credits a differing-context attribute only when both its work and personal selections are correct. PET credits a changed cell only when its pre-event and final selections are both correct. Eligible and

Table 12: Floored final probes by memory condition and backbone. Each cell has 336 final probes (4 users $\times$ 28 probes $\times$ 3 repeats). A floored probe scores PSA $= 0$ and BRS $= 1$ and stays in the denominator.

Memory	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	Total
No Memory	1	0	4	3	8
File memory	0	0	6	3	9
Mem0	0	0	8	5	13
MemOS	0	4	6	6	16
Simple RAG	0	2	10	4	16
Oracle	3	0	0	2	5
Total	4	6	34	23	67

passing counts for CS and PET are pooled across users within each repeat. We average the resulting scores over three repeats. For repeat-level values $M^{(1)}, M^{(2)}, M^{(3)}$, the reported sample standard deviation is

$$s_M = \sqrt{\frac{\sum_{r=1}^{3} \left(M^{(r)} - \overline{M}\right)^2}{2}}. \tag{3}$$

PSA, CS, and PET are reported as percentages; BRS remains on its 1–5 scale. A score with no eligible items is reported as N/A.

F.4 Cluster-Bootstrap Intervals

Table 13 gives 95% intervals for each memory condition and for its difference from No Memory. The intervals come from a cluster bootstrap with 5,000 resamples. For PSA, each resample draws the 28 final probes of each user with replacement. For CS and PET, it draws the pairs and cells with replacement. A drawn probe, pair, or cell brings its results for all four backbones and three repeats, so the three repeats of a unit always enter a resample together. Each interval spans the 2.5th to 97.5th percentiles of the resampled values.

Only File memory differs clearly from No Memory on PSA. Its mean gain of 9.6 points has a 95% interval of [+3.9, +15.6], and it beats No Memory in 14 of the 16 user–backbone pairs. Mem0, MemOS, and Simple RAG differ from No Memory by +0.4, +0.4, and +0.2 points, with 95% intervals that stay within 2.7 points of No Memory, and No Memory beats them in 10, 8, and 8 of the 16 pairs.

Carry-over counts the CS pairs in which the PA selects the same setting in both contexts, where this setting is the target of one context and differs from the backbone's default under No Memory (Finding 5). Each condition has 216 pair instances, from 18 pairs, four backbones, and three repeats. File memory has 58 carry-overs and No Memory has 13. File memory has more on every backbone, with 17 against 3 for GPT-5.6-Sol, 23 against 5 for Qwen3.8-27B, 10 against 2 for HY4-Preview, and 8 against 3 for DeepSeek-V4-Flash. It also has more for every user (5 against 1 for User A, 17 against 2 for User B, 11 against 4 for User C, and 25 against 6 for User D) and in every repeat (19, 18, and 21 against 5, 2, and 6).

Table 13: Values and 95% cluster-bootstrap intervals (%) for PSA, CS, PET, and carry-over, pooled over the four users, four backbones, and three repeats. PSA resamples the final probes within each user, and CS and PET resample their pairs and cells. Every resampled unit keeps its results for all backbones and repeats. Carry-over is the share of CS pairs in which the PA selects the same setting in both contexts, where this setting is the target of one context and differs from the backbone's default under No Memory. The right block gives the difference from No Memory. Each interval uses 5,000 resamples.

Memory	PSA	CS	PET	Carry-over	ΔPSA	ΔCS	ΔPET	ΔCarry-over
No Memory	49.6 [42.2, 57.0]	21.8 [10.2, 35.2]	11.2 [4.6, 19.6]	6.0 [1.9, 10.6]	–	–	–	–
File memory	59.2 [53.2, 65.3]	14.4 [8.3, 20.8]	25.4 [15.0, 36.7]	26.9 [17.1, 38.0]	+9.6 [+3.9, +15.6]	−7.4 [−17.1, +1.4]	+14.2 [+4.2, +25.8]	+20.8 [+10.2, +31.9]
Mem0	50.0 [42.6, 57.2]	22.2 [13.0, 32.9]	14.2 [6.7, 22.5]	4.6 [1.9, 7.9]	+0.4 [−1.5, +2.2]	+0.5 [−5.6, +6.5]	+2.9 [−0.4, +6.7]	−1.4 [−5.1, +1.4]
MemOS	50.0 [42.6, 57.3]	19.9 [10.6, 30.6]	10.8 [5.0, 17.5]	7.4 [2.8, 12.5]	+0.4 [−1.5, +2.1]	−1.9 [−9.7, +4.6]	−0.4 [−5.4, +4.2]	+1.4 [−1.9, +4.6]
Simple RAG	49.9 [42.6, 57.0]	20.8 [12.0, 31.5]	15.0 [8.3, 22.9]	6.9 [2.8, 12.0]	+0.2 [−2.2, +2.7]	−0.9 [−7.9, +5.6]	+3.8 [−0.8, +9.2]	+0.9 [−3.7, +5.6]

Table 14: Preference Selection Accuracy (PSA, %) by interaction-preference dimension, user, backbone, and memory condition.

Memory	User A				User B				User C				User D			
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash
Expression Style																
File memory	87.5 ± 0.0	83.3 ± 7.2	87.5 ± 12.5	79.2 ± 7.2	50.0 ± 0.0	75.0 ± 12.5	95.8 ± 7.2	20.8 ± 14.4	62.5 ± 12.5	79.2 ± 7.2	70.8 ± 14.4	79.2 ± 19.1	58.3 ± 7.2	70.8 ± 7.2	50.0 ± 0.0	50.0 ± 0.0
Simple RAG	50.0 ± 12.5	75.0 ± 12.5	62.5 ± 12.5	58.3 ± 14.4	54.2 ± 7.2	37.5 ± 12.5	41.7 ± 7.2	29.2 ± 19.1	50.0 ± 0.0	87.5 ± 0.0	70.8 ± 7.2	79.2 ± 7.2	37.5 ± 0.0	79.2 ± 7.2	37.5 ± 0.0	75.0 ± 12.5
Mem0	62.5 ± 0.0	79.2 ± 7.2	50.0 ± 21.7	54.2 ± 7.2	50.0 ± 0.0	41.7 ± 19.1	50.0 ± 12.5	41.7 ± 14.4	50.0 ± 0.0	79.2 ± 7.2	62.5 ± 12.5	70.8 ± 7.2	37.5 ± 0.0	83.3 ± 7.2	41.7 ± 7.2	58.3 ± 7.2
MemOS	62.5 ± 0.0	75.0 ± 0.0	70.8 ± 7.2	58.3 ± 7.2	54.2 ± 7.2	33.3 ± 14.4	41.7 ± 7.2	29.2 ± 7.2	50.0 ± 0.0	75.0 ± 12.5	54.2 ± 7.2	83.3 ± 7.2	37.5 ± 0.0	91.7 ± 14.4	41.7 ± 7.2	62.5 ± 12.5
No Memory	62.5 ± 0.0	75.0 ± 0.0	75.0 ± 0.0	62.5 ± 0.0	50.0 ± 0.0	58.3 ± 7.2	50.0 ± 0.0	29.2 ± 14.4	54.2 ± 7.2	70.8 ± 14.4	62.5 ± 12.5	70.8 ± 19.1	37.5 ± 0.0	70.8 ± 7.2	41.7 ± 7.2	54.2 ± 7.2
Oracle	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	91.7 ± 7.2	100.0 ± 0.0
Disclosure																
File memory	33.3 ± 0.0	94.4 ± 9.6	38.9 ± 25.5	66.7 ± 0.0	38.9 ± 19.2	77.8 ± 9.6	27.8 ± 9.6	83.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	50.0 ± 16.7	83.3 ± 0.0	50.0 ± 0.0	83.3 ± 0.0	83.3 ± 0.0
Simple RAG	33.3 ± 0.0	33.3 ± 0.0	38.9 ± 9.6	33.3 ± 0.0	83.3 ± 0.0	100.0 ± 0.0	77.8 ± 9.6	72.2 ± 9.6	27.8 ± 9.6	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	77.8 ± 9.6	83.3 ± 0.0	72.2 ± 19.2	77.8 ± 9.6
Mem0	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	77.8 ± 9.6	88.9 ± 9.6	88.9 ± 9.6	77.8 ± 9.6	33.3 ± 0.0	33.3 ± 0.0	27.8 ± 9.6	27.8 ± 9.6	83.3 ± 0.0	83.3 ± 16.7	83.3 ± 0.0	83.3 ± 0.0
MemOS	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	72.2 ± 9.6	83.3 ± 0.0	77.8 ± 9.6	83.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	27.8 ± 9.6	22.2 ± 9.6	83.3 ± 0.0	83.3 ± 0.0	77.8 ± 9.6	72.2 ± 19.2
No Memory	33.3 ± 0.0	27.8 ± 9.6	33.3 ± 0.0	38.9 ± 9.6	77.8 ± 9.6	83.3 ± 0.0	83.3 ± 0.0	83.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	33.3 ± 0.0	27.8 ± 9.6	77.8 ± 9.6	83.3 ± 0.0	72.2 ± 19.2	77.8 ± 9.6
Oracle	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	94.4 ± 9.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Initiative & Autonomy																
File memory	50.0 ± 0.0	56.7 ± 5.8	16.7 ± 11.5	46.7 ± 5.8	83.3 ± 11.5	86.7 ± 5.8	80.0 ± 10.0	83.3 ± 5.8	66.7 ± 5.8	36.7 ± 5.8	50.0 ± 10.0	36.7 ± 5.8	56.7 ± 5.8	60.0 ± 0.0	73.3 ± 5.8	60.0 ± 0.0
Simple RAG	20.0 ± 10.0	20.0 ± 10.0	20.0 ± 0.0	33.3 ± 5.8	56.7 ± 15.3	46.7 ± 5.8	30.0 ± 10.0	70.0 ± 0.0	46.7 ± 5.8	46.7 ± 15.3	60.0 ± 0.0	46.7 ± 5.8	46.7 ± 11.5	53.3 ± 5.8	60.0 ± 10.0	46.7 ± 5.8
Mem0	13.3 ± 5.8	26.7 ± 5.8	26.7 ± 5.8	23.3 ± 5.8	60.0 ± 0.0	50.0 ± 10.0	43.3 ± 11.5	70.0 ± 0.0	46.7 ± 15.3	50.0 ± 10.0	50.0 ± 10.0	43.3 ± 5.8	63.3 ± 5.8	56.7 ± 5.8	60.0 ± 0.0	53.3 ± 5.8
MemOS	13.3 ± 5.8	26.7 ± 5.8	16.7 ± 5.8	23.3 ± 5.8	60.0 ± 10.0	43.3 ± 5.8	56.7 ± 5.8	70.0 ± 0.0	36.7 ± 5.8	43.3 ± 5.8	56.7 ± 15.3	40.0 ± 0.0	56.7 ± 5.8	60.0 ± 10.0	66.7 ± 5.8	63.3 ± 5.8
No Memory	23.3 ± 11.5	10.0 ± 0.0	23.3 ± 5.8	26.7 ± 5.8	53.3 ± 5.8	53.3 ± 5.8	53.3 ± 5.8	70.0 ± 0.0	46.7 ± 11.5	46.7 ± 15.3	50.0 ± 10.0	53.3 ± 15.3	53.3 ± 5.8	63.3 ± 11.5	33.3 ± 5.8	56.7 ± 5.8
Oracle	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	96.7 ± 5.8	100.0 ± 0.0	100.0 ± 0.0	96.7 ± 5.8	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	96.7 ± 5.8	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
Information Flow																
File memory	75.0 ± 0.0	75.0 ± 0.0	91.7 ± 14.4	75.0 ± 25.0	58.3 ± 14.4	25.0 ± 0.0	50.0 ± 0.0	25.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	41.7 ± 14.4	8.3 ± 14.4	50.0 ± 0.0	50.0 ± 0.0	33.3 ± 14.4	50.0 ± 25.0
Simple RAG	66.7 ± 14.4	75.0 ± 0.0	75.0 ± 0.0	58.3 ± 14.4	33.3 ± 14.4	25.0 ± 0.0	41.7 ± 14.4	25.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	8.3 ± 14.4	41.7 ± 28.9	50.0 ± 25.0	41.7 ± 14.4	66.7 ± 14.4
Mem0	75.0 ± 0.0	75.0 ± 0.0	66.7 ± 14.4	50.0 ± 0.0	25.0 ± 0.0	33.3 ± 14.4	33.3 ± 14.4	25.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	58.3 ± 14.4	33.3 ± 14.4	58.3 ± 14.4	16.7 ± 14.4
MemOS	75.0 ± 0.0	75.0 ± 0.0	75.0 ± 0.0	58.3 ± 14.4	33.3 ± 14.4	25.0 ± 0.0	25.0 ± 0.0	25.0 ± 0.0	0.0 ± 0.0	8.3 ± 14.4	0.0 ± 0.0	0.0 ± 0.0	41.7 ± 14.4	50.0 ± 25.0	58.3 ± 14.4	50.0 ± 25.0
No Memory	58.3 ± 14.4	75.0 ± 0.0	75.0 ± 0.0	50.0 ± 25.0	33.3 ± 14.4	25.0 ± 0.0	25.0 ± 0.0	25.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	50.0 ± 25.0	50.0 ± 25.0	58.3 ± 14.4	41.7 ± 14.4
Oracle	100.0 ± 0.0	91.7 ± 14.4	100.0 ± 0.0	100.0 ± 0.0	91.7 ± 14.4	100.0 ± 0.0	100.0 ± 0.0	83.3 ± 14.4	91.7 ± 14.4	100.0 ± 0.0	100.0 ± 0.0	83.3 ± 14.4	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

Each cell reports dimension-specific mean ± sample SD across three repeated probe evaluations under the same configuration (PSA, %).

F.5 PSA BY INTERACTION PREFERENCE DIMENSION

Table 14 separates PSA into the four interaction-preference dimensions for each user, model, and memory condition, reporting means and sample standard deviations across three repeated probe evaluations under the same configuration.

Finding 6 groups the preferences by whether they change the PA's reply or its actions in the task. Under No Memory, the PA selects preferences on its actions correctly on 40.3% of the final probes, below the 44.4% of random selection, and preferences on its reply on 56.6%. File memory raises the two rates to 54.3% and 62.9%.

BRS and PRS for each user, backbone, and memory condition are given in Tables 15 and 16, and their values by dimension are given in Tables 17 and 18. BRS averaged over the users is shown for each backbone and memory condition in Figure 9.

Table 15: Behavioral Realization Score (BRS, 1–5) across users, backbones, and memory conditions.

Memory	User A				User B				User C				User D			
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash
File memory	3.74	4.01	3.37	4.04	3.80	3.98	3.86	3.79	2.82	2.79	3.12	2.70	3.51	3.52	3.67	3.33
Simple RAG	2.87	2.88	2.95	3.18	3.56	3.40	3.20	3.25	2.52	2.85	3.04	2.88	3.17	3.52	3.57	3.48
Mem0	2.94	3.18	2.85	3.24	3.55	3.26	3.26	3.64	2.63	2.92	2.90	2.63	3.36	3.69	3.50	3.43
MemOS	3.05	3.14	2.94	3.14	3.54	3.12	3.23	3.52	2.67	2.70	2.89	2.73	3.05	3.73	3.52	3.50
No Memory	2.87	2.96	3.04	3.21	3.43	3.62	3.13	3.49	2.64	2.60	2.85	2.93	3.15	3.83	3.02	3.52
Oracle	4.74	4.85	4.90	4.92	4.75	4.68	4.65	4.35	4.67	4.27	4.57	4.23	4.56	4.55	4.43	4.25

Each cell reports mean BRS (1–5) over three probe runs.

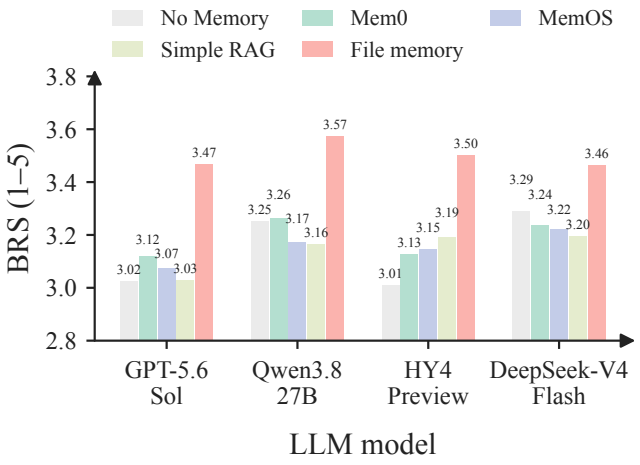


Figure 9: BRS across backbones and memory conditions, averaged over the three probe repeats and then over the four users. BRS is the LLM judge's 1–5 score of how well the delivered interaction follows the target setting, and Oracle scores are given in Table 15.

Table 16: Perceived Realization Score (PRS, 1–5) across users, backbones, and memory conditions.

Memory	User A				User B				User C				User D			
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash
File memory	3.81	4.04	3.57	4.08	3.83	3.82	3.83	3.85	3.12	2.81	3.26	2.86	3.60	3.75	3.81	3.65
Simple RAG	3.18	3.01	3.13	3.33	3.69	3.26	3.27	3.30	2.79	2.99	2.93	2.75	3.35	3.65	3.71	3.55
Mem0	3.13	3.26	3.01	3.35	3.54	3.36	3.36	3.61	2.96	3.00	3.01	2.73	3.44	3.73	3.87	3.15
MemOS	3.25	3.25	2.95	3.24	3.56	3.18	3.43	3.56	2.69	2.75	3.07	2.86	3.33	3.62	3.88	3.48
No Memory	2.96	3.04	3.23	3.29	3.52	3.65	3.29	3.42	2.99	2.94	2.90	2.94	3.30	3.94	3.55	3.52
Oracle	4.42	4.81	4.77	4.77	4.48	4.63	4.61	4.26	4.37	4.01	4.39	3.87	4.37	4.33	4.48	4.14

Each cell reports mean PRS (1–5) over three probe runs.

Table 17: Behavioral Realization Score (BRS, 1–5) by interaction-preference dimension, user, backbone, and memory condition.

Memory	User A				User B				User C				User D			
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash
Expression Style																
File memory	4.88 ± 0.22	4.54 ± 0.07	4.75 ± 0.22	4.58 ± 0.07	3.58 ± 0.52	4.12 ± 0.33	4.71 ± 0.26	3.58 ± 0.36	3.38 ± 0.43	3.96 ± 0.31	3.79 ± 0.14	3.62 ± 0.54	3.33 ± 0.38	3.58 ± 0.31	3.25 ± 0.00	2.96 ± 0.19
Simple RAG	3.46 ± 0.63	4.00 ± 0.50	3.88 ± 0.50	4.25 ± 0.33	3.42 ± 0.14	3.08 ± 0.19	3.21 ± 0.36	3.25 ± 0.45	2.92 ± 0.07	4.08 ± 0.40	4.00 ± 0.12	4.12 ± 0.12	2.92 ± 0.29	3.79 ± 0.19	3.25 ± 0.33	3.50 ± 0.22
Mem0	3.83 ± 0.26	4.46 ± 0.31	3.42 ± 0.47	4.04 ± 0.14	3.62 ± 0.25	2.75 ± 0.54	3.21 ± 0.36	3.75 ± 0.25	2.96 ± 0.19	4.04 ± 0.31	3.96 ± 0.14	3.79 ± 0.38	2.83 ± 0.07	4.17 ± 0.56	3.21 ± 0.31	3.71 ± 0.26
MemOS	4.33 ± 0.31	4.12 ± 0.12	4.21 ± 0.14	3.92 ± 0.40	3.42 ± 0.36	2.75 ± 0.43	2.96 ± 0.47	3.50 ± 0.12	3.38 ± 0.12	3.79 ± 0.44	3.42 ± 0.19	3.96 ± 0.14	2.75 ± 0.00	4.04 ± 0.44	2.96 ± 0.26	3.54 ± 0.52
No Memory	3.79 ± 0.26	4.29 ± 0.14	4.25 ± 0.22	4.17 ± 0.31	3.29 ± 0.26	3.62 ± 0.33	3.12 ± 0.22	3.29 ± 0.52	3.17 ± 0.38	3.62 ± 0.25	3.83 ± 0.26	4.04 ± 0.44	2.75 ± 0.33	4.08 ± 0.07	2.92 ± 0.44	3.75 ± 0.33
Oracle	5.00 ± 0.00	5.00 ± 0.00	4.92 ± 0.14	5.00 ± 0.00	5.00 ± 0.00	4.83 ± 0.19	4.54 ± 0.07	4.50 ± 0.12	4.92 ± 0.14	4.54 ± 0.26	4.58 ± 0.38	4.38 ± 0.43	4.33 ± 0.26	4.42 ± 0.14	4.04 ± 0.07	3.96 ± 0.14
Disclosure																
File memory	3.06 ± 0.35	4.78 ± 0.38	3.28 ± 0.67	4.00 ± 0.00	2.94 ± 0.96	3.78 ± 0.19	2.78 ± 0.69	4.22 ± 0.10	2.33 ± 0.00	2.06 ± 0.10	2.11 ± 0.10	2.67 ± 0.44	4.11 ± 0.10	3.67 ± 0.17	4.06 ± 0.35	4.11 ± 0.19
Simple RAG	2.72 ± 0.10	2.67 ± 0.00	2.83 ± 0.29	2.89 ± 0.38	4.33 ± 0.00	4.44 ± 0.19	4.06 ± 0.35	3.78 ± 0.35	2.17 ± 0.29	2.06 ± 0.10	2.06 ± 0.10	2.39 ± 0.42	4.11 ± 0.25	3.94 ± 0.25	3.94 ± 0.38	4.28 ± 0.10
Mem0	3.11 ± 0.38	2.78 ± 0.19	2.61 ± 0.10	3.11 ± 0.38	4.11 ± 0.38	4.33 ± 0.33	4.17 ± 0.29	4.17 ± 0.29	2.50 ± 0.29	2.06 ± 0.10	2.06 ± 0.19	2.00 ± 0.29	4.06 ± 0.19	4.06 ± 0.19	4.17 ± 0.17	4.33 ± 0.00
MemOS	2.89 ± 0.38	2.78 ± 0.19	2.83 ± 0.44	2.89 ± 0.38	3.72 ± 0.59	4.00 ± 0.29	3.89 ± 0.48	4.28 ± 0.10	2.50 ± 0.29	2.00 ± 0.00	2.22 ± 0.42	1.94 ± 0.19	4.28 ± 0.10	4.11 ± 0.10	4.17 ± 0.17	3.78 ± 0.67
No Memory	2.72 ± 0.10	2.61 ± 0.25	2.67 ± 0.00	3.06 ± 0.35	4.06 ± 0.48	4.22 ± 0.10	4.17 ± 0.17	4.33 ± 0.00	2.56 ± 0.25	2.11 ± 0.10	2.06 ± 0.10	2.22 ± 0.35	4.11 ± 0.10	4.00 ± 0.17	3.72 ± 0.77	3.89 ± 0.35
Oracle	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.89 ± 0.10	3.94 ± 0.19	4.56 ± 0.35	4.61 ± 0.35	5.00 ± 0.00	4.06 ± 0.35	4.06 ± 0.25	3.94 ± 0.35	4.78 ± 0.25	4.22 ± 0.10	4.50 ± 0.00	4.22 ± 0.25
Initiative & Autonomy																
File memory	3.27 ± 0.06	3.40 ± 0.10	2.00 ± 0.60	3.70 ± 0.17	4.77 ± 0.32	4.43 ± 0.29	4.23 ± 0.67	4.23 ± 0.31	3.27 ± 0.12	2.87 ± 0.42	3.33 ± 0.59	2.53 ± 0.15	3.43 ± 0.70	3.30 ± 0.30	4.03 ± 0.23	3.10 ± 0.26
Simple RAG	2.20 ± 0.44	1.97 ± 0.35	2.00 ± 0.10	2.53 ± 0.21	3.63 ± 0.67	3.53 ± 0.40	2.63 ± 0.45	3.40 ± 0.17	2.97 ± 0.06	3.03 ± 0.29	3.60 ± 0.10	2.77 ± 0.21	2.90 ± 0.46	3.17 ± 0.21	3.77 ± 0.23	2.90 ± 0.70
Mem0	1.90 ± 0.10	2.30 ± 0.26	2.20 ± 0.10	2.60 ± 0.61	3.60 ± 0.36	3.30 ± 0.26	2.80 ± 0.44	3.80 ± 0.00	2.97 ± 0.32	3.23 ± 0.06	3.17 ± 0.32	2.67 ± 0.23	3.50 ± 0.00	3.40 ± 0.53	3.30 ± 0.20	3.10 ± 0.56
MemOS	1.83 ± 0.15	2.37 ± 0.06	1.70 ± 0.17	2.57 ± 0.55	3.83 ± 0.25	3.17 ± 0.06	3.30 ± 0.26	3.50 ± 0.00	2.70 ± 0.26	2.80 ± 0.36	3.60 ± 0.46	2.90 ± 0.52	3.00 ± 0.53	3.50 ± 0.10	3.60 ± 0.56	3.47 ± 0.74
No Memory	2.27 ± 0.12	1.80 ± 0.35	2.13 ± 0.12	2.53 ± 0.25	3.47 ± 0.06	3.53 ± 0.35	2.90 ± 0.17	3.53 ± 0.47	2.90 ± 0.44	2.67 ± 0.45	3.17 ± 0.46	3.23 ± 0.35	3.17 ± 0.23	3.73 ± 0.12	2.43 ± 0.42	3.20 ± 0.20
Oracle	4.93 ± 0.12	4.87 ± 0.15	4.80 ± 0.20	4.80 ± 0.17	4.73 ± 0.12	4.93 ± 0.06	4.70 ± 0.20	4.03 ± 0.32	4.63 ± 0.15	4.30 ± 0.30	4.83 ± 0.21	4.27 ± 0.15	4.63 ± 0.06	4.97 ± 0.06	4.67 ± 0.25	4.33 ± 0.06
Information Flow																
File memory	3.67 ± 0.14	3.33 ± 0.14	4.17 ± 0.14	3.83 ± 1.04	3.08 ± 0.38	2.83 ± 0.38	2.83 ± 0.38	2.42 ± 0.38	1.33 ± 0.58	1.33 ± 0.58	2.75 ± 0.90	1.33 ± 0.58	3.17 ± 0.88	3.75 ± 0.43	3.00 ± 0.43	3.50 ± 0.50
Simple RAG	3.58 ± 0.52	3.25 ± 0.25	3.67 ± 0.38	3.08 ± 0.14	2.50 ± 0.25	2.17 ± 0.14	3.33 ± 0.38	2.08 ± 0.14	1.17 ± 0.14	1.08 ± 0.14	1.17 ± 0.14	1.42 ± 0.52	2.92 ± 0.14	3.25 ± 0.90	3.17 ± 0.58	3.67 ± 0.29
Mem0	3.50 ± 0.25	3.42 ± 0.29	3.67 ± 0.38	3.42 ± 0.52	2.42 ± 0.52	2.58 ± 0.38	3.17 ± 0.14	2.25 ± 0.43	1.33 ± 0.58	1.17 ± 0.29	1.42 ± 0.29	1.17 ± 0.29	3.00 ± 0.50	2.92 ± 1.28	3.58 ± 0.38	2.33 ± 0.29
MemOS	3.75 ± 0.25	3.67 ± 0.29	3.67 ± 0.38	3.42 ± 0.38	2.75 ± 0.00	2.42 ± 0.14	2.58 ± 0.63	2.50 ± 0.25	1.42 ± 0.38	1.33 ± 0.38	1.08 ± 0.14	1.00 ± 0.00	1.92 ± 0.76	3.08 ± 0.72	3.50 ± 0.25	3.08 ± 0.38
No Memory	2.75 ± 0.43	3.75 ± 0.25	3.42 ± 0.14	3.25 ± 0.66	2.67 ± 0.14	2.92 ± 0.63	2.17 ± 0.14	2.50 ± 0.66	1.08 ± 0.14	1.08 ± 0.14	1.25 ± 0.25	1.00 ± 0.00	2.50 ± 0.25	3.33 ± 0.14	3.67 ± 0.29	3.33 ± 0.14
Oracle	3.33 ± 0.52	4.25 ± 0.66	5.00 ± 0.00	4.92 ± 0.14	4.08 ± 0.52	4.83 ± 0.29	4.92 ± 0.14	4.42 ± 0.52	3.75 ± 0.90	4.00 ± 0.50	4.67 ± 0.38	4.25 ± 0.75	4.50 ± 0.00	4.25 ± 0.25	4.50 ± 0.43	4.67 ± 0.29

Each cell reports dimension-specific mean ± sample SD across three repeated probe evaluations under the same configuration (BRS, 1–5).

Table 18: Perceived Realization Score (PRS, 1–5) by interaction-preference dimension, user, backbone, and memory condition.

Memory	User A				User B				User C				User D			
	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash	GPT-5.6-Sol	Qwen3.8-27B	HY4-Preview	DeepSeek-V4-Flash
Expression Style																
File memory	4.88 ± 0.22	4.33 ± 0.19	4.75 ± 0.22	4.58 ± 0.07	4.08 ± 0.52	3.75 ± 0.25	4.71 ± 0.29	3.62 ± 0.45	3.54 ± 0.29	3.75 ± 0.33	4.00 ± 0.33	3.50 ± 0.22	4.04 ± 0.19	4.25 ± 0.00	4.17 ± 0.07	3.50 ± 0.22
Simple RAG	3.54 ± 0.52	4.08 ± 0.51	3.96 ± 0.26	3.62 ± 0.45	3.92 ± 0.14	3.04 ± 0.29	3.25 ± 0.25	3.04 ± 0.26	3.29 ± 0.07	3.92 ± 0.19	3.88 ± 0.33	3.46 ± 0.40	3.29 ± 0.14	4.12 ± 0.25	3.83 ± 0.07	3.92 ± 0.26
Mem0	3.75 ± 0.12	4.33 ± 0.26	3.42 ± 0.29	3.46 ± 0.19	3.62 ± 0.12	2.88 ± 0.33	3.42 ± 0.29	3.83 ± 0.14	3.50 ± 0.22	3.83 ± 0.38	3.92 ± 0.19	3.46 ± 0.19	3.38 ± 0.12	4.21 ± 0.31	3.88 ± 0.12	3.75 ± 0.38
MemOS	4.04 ± 0.14	4.04 ± 0.26	4.17 ± 0.14	3.62 ± 0.25	3.62 ± 0.33	2.79 ± 0.31	3.33 ± 0.26	3.67 ± 0.07	3.38 ± 0.12	3.71 ± 0.31	3.88 ± 0.22	3.71 ± 0.19	3.46 ± 0.19	4.21 ± 0.14	3.88 ± 0.12	3.88 ± 0.38
No Memory	3.83 ± 0.07	4.12 ± 0.12	4.21 ± 0.07	3.92 ± 0.19	3.67 ± 0.26	3.67 ± 0.44	3.42 ± 0.07	3.08 ± 0.62	3.67 ± 0.26	3.75 ± 0.22	4.08 ± 0.31	3.67 ± 0.14	3.21 ± 0.07	4.17 ± 0.14	3.75 ± 0.25	3.71 ± 0.19
Oracle	4.96 ± 0.07	5.00 ± 0.00	4.67 ± 0.29	4.92 ± 0.07	4.83 ± 0.19	4.75 ± 0.12	4.62 ± 0.12	4.25 ± 0.33	4.50 ± 0.22	4.21 ± 0.07	4.33 ± 0.19	3.83 ± 0.26	4.67 ± 0.19	4.42 ± 0.29	4.29 ± 0.07	4.25 ± 0.54
Disclosure																
File memory	3.44 ± 0.35	4.67 ± 0.33	3.39 ± 1.02	4.06 ± 0.10	2.72 ± 0.48	3.44 ± 0.10	2.33 ± 0.44	4.17 ± 0.00	2.56 ± 0.10	1.72 ± 0.25	1.94 ± 0.51	2.39 ± 0.67	3.94 ± 0.25	3.28 ± 0.42	3.50 ± 0.44	3.94 ± 0.42
Simple RAG	3.17 ± 0.29	2.78 ± 0.19	3.00 ± 0.29	3.06 ± 0.59	4.11 ± 0.25	3.89 ± 0.25	4.06 ± 0.42	3.61 ± 0.19	2.39 ± 0.48	2.33 ± 0.17	1.89 ± 0.10	1.89 ± 0.38	4.06 ± 0.19	3.83 ± 0.29	3.78 ± 0.51	4.11 ± 0.54
Mem0	3.33 ± 0.50	3.17 ± 0.00	3.33 ± 0.29	3.44 ± 0.35	4.11 ± 0.25	4.22 ± 0.19	3.83 ± 0.29	3.61 ± 0.59	2.56 ± 0.10	2.22 ± 0.19	2.06 ± 0.42	2.11 ± 0.48	3.94 ± 0.25	3.83 ± 0.29	3.89 ± 0.25	3.39 ± 0.63
MemOS	3.33 ± 0.17	2.94 ± 0.38	2.83 ± 0.17	3.17 ± 0.17	3.83 ± 0.50	3.94 ± 0.42	3.78 ± 0.35	3.72 ± 0.25	2.33 ± 0.17	1.56 ± 0.25	2.17 ± 0.44	2.17 ± 0.29	3.94 ± 0.42	3.56 ± 0.35	4.17 ± 0.29	3.56 ± 0.48
No Memory	3.11 ± 0.63	2.89 ± 0.35	2.89 ± 0.25	3.17 ± 0.60	3.72 ± 0.69	4.06 ± 0.10	3.72 ± 0.38	3.83 ± 0.44	2.67 ± 0.17	2.06 ± 0.10	1.83 ± 0.29	2.11 ± 0.59	4.00 ± 0.17	3.72 ± 0.10	3.67 ± 0.60	3.67 ± 0.29
Oracle	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	5.00 ± 0.00	4.33 ± 0.44	3.89 ± 0.25	4.17 ± 0.17	4.00 ± 0.44	4.61 ± 0.38	3.78 ± 0.10	3.89 ± 0.25	3.50 ± 0.44	4.39 ± 0.38	3.83 ± 0.33	4.11 ± 0.10	3.78 ± 0.19
Initiative & Autonomy																
File memory	3.30 ± 0.10	3.70 ± 0.10	2.57 ± 0.50	3.93 ± 0.23	4.57 ± 0.51	4.57 ± 0.23	4.27 ± 0.50	4.33 ± 0.15	3.57 ± 0.15	2.97 ± 0.15	3.67 ± 0.31	2.97 ± 0.23	3.33 ± 0.21	3.57 ± 0.15	4.03 ± 0.35	3.63 ± 0.31
Simple RAG	2.70 ± 0.26	2.17 ± 0.25	2.40 ± 0.36	3.30 ± 0.26	3.60 ± 0.56	3.27 ± 0.21	2.67 ± 0.55	3.60 ± 0.17	3.00 ± 0.00	3.17 ± 0.15	3.40 ± 0.10	3.10 ± 0.10	3.03 ± 0.23	3.27 ± 0.21	3.90 ± 0.26	2.87 ± 0.32
Mem0	2.40 ± 0.26	2.50 ± 0.00	2.30 ± 0.17	3.17 ± 0.31	3.50 ± 0.10	3.33 ± 0.21	2.97 ± 0.49	3.73 ± 0.12	3.13 ± 0.38	3.33 ± 0.31	3.40 ± 0.30	3.03 ± 0.25	3.27 ± 0.15	3.53 ± 0.29	3.93 ± 0.15	2.83 ± 0.32
MemOS	2.47 ± 0.32	2.70 ± 0.20	1.93 ± 0.42	3.00 ± 0.46	3.57 ± 0.35	3.20 ± 0.00	3.60 ± 0.20	3.73 ± 0.21	2.80 ± 0.26	3.03 ± 0.31	3.57 ± 0.60	3.17 ± 0.12	3.37 ± 0.06	3.37 ± 0.29	3.80 ± 0.35	3.17 ± 0.40
No Memory	2.43 ± 0.25	2.13 ± 0.25	2.50 ± 0.00	2.93 ± 0.06	3.53 ± 0.15	3.57 ± 0.21	3.17 ± 0.29	3.67 ± 0.15	3.13 ± 0.25	3.37 ± 0.50	3.03 ± 0.35	3.37 ± 0.12	3.13 ± 0.35	3.93 ± 0.47	3.23 ± 0.15	3.23 ± 0.15
Oracle	4.37 ± 0.21	4.70 ± 0.17	4.67 ± 0.25	4.60 ± 0.36	4.60 ± 0.10	4.97 ± 0.06	4.73 ± 0.15	4.43 ± 0.21	4.53 ± 0.21	4.10 ± 0.20	4.63 ± 0.12	4.17 ± 0.31	4.17 ± 0.21	4.60 ± 0.17	4.87 ± 0.23	4.13 ± 0.42
Information Flow																
File memory	3.50 ± 0.00	3.33 ± 0.29	4.00 ± 0.25	3.50 ± 0.25	3.17 ± 0.52	2.67 ± 0.52	3.25 ± 0.00	2.58 ± 0.14	2.00 ± 0.66	2.17 ± 0.29	2.75 ± 0.25	2.00 ± 0.43	2.83 ± 0.38	3.92 ± 0.58	3.00 ± 0.43	3.58 ± 0.38
Simple RAG	3.67 ± 0.29	3.33 ± 0.14	3.50 ± 0.25	3.25 ± 0.25	2.83 ± 0.38	2.75 ± 0.00	3.67 ± 0.29	2.58 ± 0.29	1.83 ± 0.14	1.67 ± 0.14	1.42 ± 0.14	1.75 ± 0.87	3.17 ± 0.38	3.42 ± 0.63	2.92 ± 0.52	3.67 ± 0.29
Mem0	3.42 ± 0.14	3.17 ± 0.14	3.50 ± 0.43	3.42 ± 0.52	2.58 ± 0.14	3.08 ± 0.52	3.50 ± 0.00	2.83 ± 0.58	2.08 ± 0.58	1.67 ± 0.14	1.67 ± 0.29	1.42 ± 0.14	3.25 ± 0.43	3.08 ± 1.18	3.67 ± 0.52	2.42 ± 0.38
MemOS	3.50 ± 0.25	3.50 ± 0.25	3.25 ± 0.00	3.17 ± 0.14	3.00 ± 0.25	2.75 ± 0.00	2.67 ± 0.14	2.67 ± 0.14	1.58 ± 0.14	1.92 ± 0.29	1.58 ± 0.14	1.42 ± 0.38	2.08 ± 0.58	3.17 ± 0.58	3.67 ± 0.29	3.33 ± 0.29
No Memory	2.33 ± 0.14	3.33 ± 0.14	3.58 ± 0.58	3.08 ± 0.88	2.92 ± 0.29	3.25 ± 0.50	2.67 ± 0.14	2.83 ± 0.58	1.75 ± 0.25	1.58 ± 0.14	1.83 ± 0.14	1.67 ± 0.14	2.83 ± 0.38	3.83 ± 0.29	3.75 ± 0.43	3.67 ± 0.29
Oracle	2.58 ± 0.63	4.42 ± 0.29	4.92 ± 0.14	4.58 ± 0.14	3.67 ± 0.38	4.67 ± 0.38	4.92 ± 0.14	4.25 ± 0.50	3.33 ± 0.95	3.75 ± 0.66	4.67 ± 0.58	3.75 ± 0.66	4.25 ± 0.00	4.25 ± 0.25	4.42 ± 0.63	4.50 ± 0.43

Each cell reports dimension-specific mean ± sample SD across three repeated probe evaluations under the same configuration (PRS, 1–5).

Table 19: PSA and general capability of the four PA backbones. Intelligence Index and GDPval-AA scores are taken from Artificial Analysis (Artificial Analysis, 2026) at the default reasoning effort of each backbone. PSA is the mean over Users A–D and the four memory conditions.

Model (default reasoning effort)	Intelligence Index	GDPval-AA (Elo)	PSA (%)
GPT-5.6-Sol (medium)	39	1403	49.8
Qwen3.8-27B (xhigh)	34	1409	56.0
HY4-Preview (high)	–	–	51.3
DeepSeek-V4-Flash (high)	26	–	52.0

Scores from Intelligence Index v4.3.2 and GDPval-AA v2.1, accessed 25 September 2026. Default reasoning efforts follow the OpenRouter model catalog on the same date. Dashes mark scores that Artificial Analysis does not list: HY4-Preview has no entry, and GDPval-AA lists DeepSeek-V4-Flash only at maximum effort. DeepSeek-V4-Flash refers to the April 2026 checkpoint, to which the OpenRouter catalog mapped this model name during our runs.

F.6 PSA and General Capability

The PSA ranking of the four backbones does not follow their general capability. Table 19 compares PSA with two scores from Artificial Analysis (Artificial Analysis, 2026). The Intelligence Index combines ten evaluations, and GDPval-AA is an Elo rating on agentic real-world work tasks. PA calls do not set a reasoning effort, so each backbone runs at its default effort, and each score comes from the matching leaderboard entry. GPT-5.6-Sol has the highest Intelligence Index of the listed backbones and the lowest PSA. Qwen3.8-27B reaches the highest PSA, and its GDPval-AA rating is close to that of GPT-5.6-Sol. Four backbones are too few for a correlation analysis, so only the rankings are compared.

F.7 Backbone Default Settings

Under No Memory the PA has no evidence about the user, and its selections show the default setting of each backbone. Table 20 gives, for each backbone and attribute, the setting that the PA selected most often under No Memory and its share of the selections.

F.8 Delivered Behavior after a Correct Selection

Under Oracle, the five settings with the lowest pass rates are Hide for Reasoning Visibility (38%), Terse for Verbosity (50%), Iterative (61%) and Structured (74%) for Information Elicitation, and Low for Solution Breadth (77%) (Finding 7). Figure 10 shows one of these shortfalls. In an Oracle final probe for User C at work, the HY4-Preview PA selects Hide for Reasoning Visibility, and the IX Wrapper withholds the rationale written in a separate field. The PA still writes three reasons in the reply under "Why that day:", and the judge gives a BRS of 1. The rest of this section gives one example per dimension in which the PA selects the target setting and the delivered reply still falls short. The judge gives each of them a BRS of 3 or lower. Quoted replies are shortened with ellipses.

Expression Style. The example targets Terse for Verbosity, which asks for the answer in the first sentence and at most one supporting move. Under File memory, User C asks the Qwen3.8-27B PA what corporate requires before an equipment purchase above $600 and which records the branch must keep. The PA selects Terse and plans "one compact block, with at most one supporting note." Its reply opens with when the rules were issued and whom they cover, lists six numbered steps, and closes with "One thing that bites people most". The judge gives a BRS of 1, and the user answers, "That is a lot of words for a simple policy."

Disclosure. The example targets Hide for Reasoning Visibility, which asks for the answer with no grounds anywhere in the reply. Under Oracle, User B asks the DeepSeek-V4-Flash PA for the latest day to collect prints from a print shop. The PA selects Hide, and its reply opens with "Here's the breakdown" and works through the collection rule and the Sunday closure before it names Saturday, February 6. The judge gives a BRS of 1.

Table 20: Default setting of each backbone under No Memory. Each cell gives the setting that the PA selects most often for the attribute, with its share and count of the selections, pooled over the final and pre-event probes of the four users and three repeats. Probes that received no score are excluded.

Attribute	**GPT-5.6-Sol**	**Qwen3.8-27B**	**HY4-Preview**	**DeepSeek-V4-Flash**
Expression Style				
Tone & Formality	Consultative 70% (23/33)	Consultative 42% (14/33)	Consultative 67% (22/33)	Consultative 91% (30/33)
Verbosity	Detailed 100% (24/24)	Detailed 71% (17/24)	Detailed 96% (23/24)	Detailed 62% (15/24)
Emotional Engagement	Relationship-focused 100% (27/27)	Relationship-focused 100% (27/27)	Relationship-focused 100% (27/27)	Relationship-focused 100% (27/27)
Guidance Level	Guided 89% (24/27)	Guided 59% (16/27)	Guided 67% (18/27)	Assumed 78% (21/27)
Disclosure				
Reasoning Visibility	Show 100% (24/24)	Show 96% (23/24)	Show 96% (23/24)	Show 100% (24/24)
Uncertainty Expression	Express 100% (27/27)	Express 100% (27/27)	Express 96% (26/27)	Express 85% (23/27)
Process Visibility	Bookend 91% (30/33)	Bookend 100% (33/33)	Bookend 94% (31/33)	Bookend 91% (30/33)
Initiative & Autonomy				
Autonomy Level	Autonomous 51% (18/35)	Autonomous 67% (24/36)	Self-directed 50% (18/36)	Self-directed 100% (36/36)
Proactive Outreach	High 88% (21/24)	High 88% (21/24)	High 82% (18/22)	High 88% (21/24)
Task Expansion	Low 100% (24/24)	Low 100% (24/24)	Low 83% (20/24)	Low 96% (23/24)
Solution Breadth	High 55% (18/33)	Low 94% (31/33)	Low 94% (31/33)	Low 97% (32/33)
Capability Boundary	Suggest alternatives 71% (17/24)	Suggest alternatives 88% (21/24)	Suggest alternatives 62% (15/24)	Suggest alternatives 79% (19/24)
Information Flow				
Information Elicitation	Structured 50% (15/30)	Infer 50% (15/30)	Infer 57% (16/28)	Infer 52% (14/27)
Topic Management	Organize 97% (28/29)	Organize 97% (29/30)	Organize 100% (30/30)	Organize 93% (28/30)

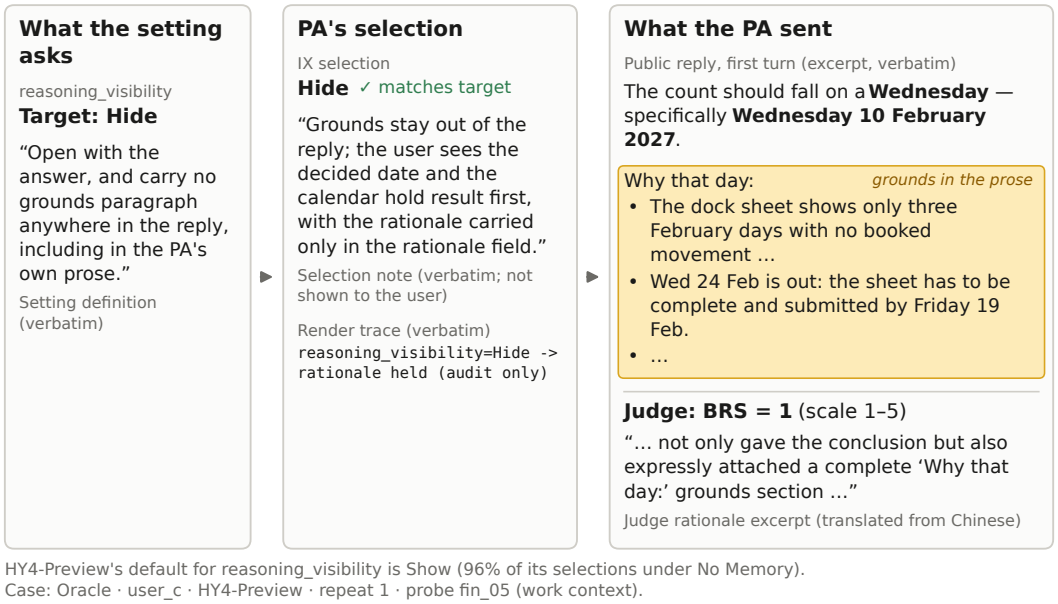


Figure 10: A correct selection whose reply still follows the backbone's default (Oracle, User C, HY4-Preview, work context; the judge's rationale is translated from Chinese). The PA selects the target Hide for Reasoning Visibility and the IX Wrapper holds back the rationale field, but the reply still adds a "Why that day:" section with the grounds, and the judge gives a BRS of 1.

Initiative & Autonomy. Under Oracle, User B asks the HY4-Preview PA which Halden desk lamp to buy for recording self-tapes. The target Low for Solution Breadth asks for one option with the reasons it works, and naming an alternative, even to show why the choice wins, counts as a list. The PA selects Low and recommends the Halden LED Desk Lamp, but it gives each reason as a comparison with the Mini, as in "Five-step dimmer vs. the Mini's two steps". The judge gives a BRS of 1, and the user answers, "you don't need to walk me through everything the other lamp couldn't do." HY4-Preview also selects Low under No Memory (Table 20), so the shortfall lies in its habit of arguing for a choice by comparing each point with the alternative.

Information Flow. The example targets Iterative for Information Elicitation, which asks the PA to request each missing fact in its own turn before it acts. User D wants a calendar hold for dropping off prints at a framing shop and has not yet chosen the day or the arrival time. Under Oracle, the HY4-Preview PA selects Iterative, but in the same call it states that "Nothing is missing this turn" and that it will make the calendar hold now. It sets the hold from the user's 17:00 departure and tells the user, "Two things I filled in myself, since you didn't specify". The user answers, "You really should have asked me what time I wanted to arrive instead of just guessing." The judge gives a BRS of 3.

F.9 Validity Checks

Is each preference exposed? Two annotators read a random 30% of each user's accumulation sessions, 28 of the 94 per user and 112 in total, all from the runs with File memory. For each session, they judged whether a reader who sees only the visible conversation can identify the target setting from the user's reactions. The preference was conveyed in 99.3% of the sessions, with an inter-annotator agreement of 95.4% (Gwet's AC1; Gwet, 2008).

Does the simulator leak a preference? The post-hoc audit in Appendix D.4 found no leakage, and we reviewed by hand every sentence that its reader marked as possible leakage.

Does the simulator follow the script? Two annotators read a random 10% of the accumulation sessions and checked that the simulator played each session as scripted: it gives each task fact at its

beat, plays the decision point and the reaction beat, and stays in character. They found no session that deviates from its script, with an inter-annotator agreement of 99.1% (Gwet's AC1).

Do BRS and PRS agree? On the 7,996 final probes that have both scores, BRS and PRS agree exactly on 65.3% and within one point on 90.0% (Spearman $\rho = 0.83$, 95% CI [0.82, 0.84]). Agreement is similar under No Memory ($\rho = 0.83$) and the memory conditions ($\rho = 0.83$) and lower under Oracle ($\rho = 0.66$), where most scores lie at 4–5. Both scores rank File memory first for every backbone (Figures 2 and 9). We also re-judged a stratified sample of 418 probes with Claude Sonnet 5, which serves neither as PA nor as simulator. Its agreement with PRS is lower than that of the Gemini judge ($\rho = 0.71$ vs. 0.85; difference -0.14, 95% CI $[-0.20, -0.09]$) but close to its agreement with the Gemini judge ($\rho = 0.76$). The BRS–PRS agreement therefore mainly reflects the interaction itself.

Which score compares settings? PRS is the user simulator's own rating, given as the user, so it carries the user's temperament along with the quality of the reply. BRS applies the same rubric for every user. The two scores for each user on the correctly selected Oracle final probes are given in Table 21. They agree on 77.3% of these probes, and PRS falls below BRS on 16.1%. This share differs by user, from 11.0% for User A to 21.5% for User C. The two scores therefore serve different purposes. PRS measures how each user receives the reply, and BRS compares settings on equal terms, since each target setting is held by a different set of users (Finding 7).

Table 21: BRS and PRS on the correctly selected Oracle final probes, by user. PRS falls below BRS more often for some users than for others, so PRS carries each user's subjective judgment and BRS serves to compare settings across users.

User	Probes	Mean BRS	Mean PRS	PRS $<$ BRS	PRS $=$ BRS	PRS $>$ BRS
User A	335	4.85	4.69	11.0%	87.2%	1.8%
User B	330	4.65	4.54	14.8%	77.9%	7.3%
User C	331	4.48	4.19	21.5%	71.6%	6.9%
User D	334	4.47	4.34	17.1%	72.5%	10.5%
All	1,330	4.61	4.44	16.1%	77.3%	6.6%

G Prompt Directory and Templates

G.1 Prompt Directory

This appendix gives the prompts of the LLM calls in the construction pipeline, at runtime, and in evaluation. Table 22 lists them by stage. Each box is named by the component that sends the prompt. The construction prompts and the two message-generation prompts of the user simulator are long, so their boxes summarize each prompt by its purpose, when it is called, its inputs and output, and its key rules, and then quote one or two passages of the original text. The other boxes show the prompts themselves, and the parts of such a box are the messages of one call or the variants of one prompt. Text in braces or angle brackets is filled in at run time. Highlighted lines are our notes or mark omitted passages, and they are not sent to the model. Non-ASCII characters are shown in ASCII, for example an em dash as `--`. The full prompts are released with the code.

G.2 Construction Prompts

The construction prompts build each user from the transcripts (Appendix B) and each session from the user's world (Appendix C). The skills with which coding agents author the worlds and check the sessions are released with the code and are not shown here. The conditions these checks apply are given in Table 8.

Table 22: Directory of the prompts in Appendices G.2–G.4. Each prompt is named by the component that sends it.

Prompt	Called	Main inputs	Output	Page
Construction				
Preference Extractor, Passes 1 and 2	Pass 1 once per scene of the source character, Pass 2 once per attribute–context cell	Scene transcript; Pass-1 evidence, attribute definition, settings, rubric	Evidence items per attribute; setting, confidence, cited evidence	49
Identity Extractor, Identity Synthesizer, Personality Rater	Once per scene, once per section group, once per character	Scene transcript; observations of all scenes	Identity sections with cited scenes; Big Five facet and Jungian function levels	50
Anonymizer and Anonymization Validator	Once per named life block; once per probe reply	Named life block; probe message, persona reply	Substitution table; attributes the persona pushed back on	51
Self-View Extractor	Once per file or cluster of files	Identity, world canon, file content	What the user recalls of the file	52
Session Outline and Accumulation Session Generators	Once per batch of sessions; once per session	Dealt slots, persona, preference definitions; outline entry, world records, tools	Session outlines; session draft with task facts and beats	52
Probe Outline and Probe Session Generators	Once per probe each	Cell, errand, attribute definition; probe skeleton, outline entry	World binding, user ask, viable answers; runtime prose of the probe	53
Change-Event Generator and Transcript Writer	Once per change event each	Old and new settings, outline entry; change-event draft	Change-event draft; fixed transcript	54
Dialogue Naturalness	Appended to four generators and the simulator	None	Constraints on user wording	55
Runtime				
User Simulator: Profile Context	Every simulator turn	Anonymized identity, tool policy, own memory	Profile block of the simulator prompt	56
User Simulator: Message Generation in Accumulation Sessions	Every simulator turn of an accumulation session	Profile, active skills, session script, current beat, task facts, dialogue	Next user message	59
User Simulator: Message Generation in Probes	Every simulator turn of a probe	Same inputs as in accumulation sessions	Next user message	60
Simulator Message Checker	Every candidate simulator message	Candidate message, PA's previous message, current beat	Verdict on each rule question	61
PA: System Prompt	Every PA call	Memory content of the condition	System message	73
PA: Interaction Tools	At decision points	Settings and their instructions	Selected setting	75
Evaluation				
Judge: System Prompt	Once per probe	Universal anchors	Scoring rules	82
Judge: User Message	Once per probe	Probe, scene, dialogue, tool trajectory, rubric	BRS score with evidence	83
Judge: Attribute Rubrics	Once per probe	Probed attribute	Criteria for the score	86

Preference Extractor: Scene Evidence and Setting Synthesis

Pass 1: Scene Evidence

Purpose. Tags the evidence one transcript scene gives about the interaction preferences of the source character. Pass 2 turns this evidence into the preference settings behind User A–D.

Called. Once per scene in which the source character speaks. Scenes with fewer than 3 dialogue turns are skipped.

Inputs.
- The 14 attributes with their settings, and the work and personal contexts
- The name of the source character, the scene reference and the scene description
- The scene transcript, one line per turn

Output. Per scene: relevance, one context, and for each tagged attribute the evidence type, a direct quote and the setting it points to.

Key rules.
- Tag an attribute only on clear behavioral evidence. Quote the dialogue directly.
- Assign each relevant scene to one context, work or personal. A scene that fits neither is marked not relevant.
- Evidence is production (how the character speaks and behaves), reception (how the character reacts to the styles of others) or an explicit statement.
- Production is only a prior. Reception is the strongest signal. When the two diverge, the extractor notes the asymmetry, and the PA preference follows reception.

Pass 2: Setting Synthesis

Purpose. Names the setting of the source character for one attribute in one context. Each result fills one cell of the 2×14 preference matrix.

Called. Once per (attribute, context) cell that has Pass-1 evidence.

Inputs.
- The attribute, its description and its settings
- The context, its description and the rubric for the attribute
- The Pass-1 evidence for the cell

Output. One of the listed settings, a confidence level, an evidence summary, and the production and reception quotes behind the setting.

Key rules.
- Three layers decide the setting. Frequency patterns give the direction, production patterns anchor a setting, and reception confirms or overrides it.
- Valid reception signals are explicit confusion or discomfort with a register, a change in engagement in response to style, verbal meta-commentary, and accommodation shifts. Plot-driven topic changes, content disagreement and scripted dramatic reactions are invalid.
- Confidence levels. High: production and reception agree across several scenes. Medium-High: strong production and some reception, or strong reception in few scenes. Medium: production only, or ambiguous reception. Low: thin or contradictory evidence.
- The setting is always one of the listed settings. When the evidence is thin or contradictory, the call still names the best-fitting setting at Low confidence. Every cell below High goes to human review.

From the Pass 1 system prompt

```
If a scene does not fit either of these 2 contexts, mark it as NOT relevant. Do not
force a classification.
```

From the Pass 2 system prompt

```
A character's speaking style is only a PRIOR. How they REACT to others' communication
styles CONFIRMS or OVERRIDES that prior. When production and reception diverge, the
preference follows RECEPTION.
```

Identity Extractor, Identity Synthesizer, and Personality Rater

Identity Extractor

Purpose. Records what one scene shows about who the source character is and how they think, talk and react. The observations feed the user identity that the simulator plays.

Called. Once per scene in which the source character speaks.

Inputs.
- The observation sections, from life facts to personality evidence
- The name of the source character, the scene reference and the scene description
- The scene transcript

Output. Observations, each with one section, a one-sentence note and a verbatim quote.

Key rules.
- Describe behavior only. The word "prefers" and the preference setting names never appear.
- Record only what this scene shows, with no outside knowledge of the show. A description by another character counts only when the behavior of the source character in the scene confirms it.
- Every observation carries a verbatim quote of at most 40 words. Observations with no quote are discarded.

Identity Synthesizer

Purpose. Merges the per-scene observations into a few stable items per section of the user identity.

Called. Once per section group (eight groups). A section with over 250 observations gets two half calls and a merge call.

Inputs.
- The sections of the call and their item caps
- Their observations, each with scene reference, note and quote
- In a merge call, the items of both halves

Output. Pattern items with their scene references. A script then sets confidence from scene and episode counts. Medium items go to human review.

Key rules.
- Merge observations that describe the same pattern and keep all their scene references. Related but distinct patterns stay apart.
- Cite only scene references from the input. Confidence is left to a script.
- Item text holds no quote, name, place or institution. Catchphrases are the one exception.

Personality Rater

Purpose. Rates the 15 Big Five facets and the 8 Jungian functions of the source character. A script derives the five dimensions and the function stack.

Called. Once per source character, plus at most one retry for ratings that carry a level with no usable evidence.

Inputs.
- The rubric with the level scale and the facet and function indicators
- The Big Five and Jungian observations, grouped by facet and function
- On a retry, the ratings to redo

Output. A level from `very_low` to `very_high`, or `null`, for each facet and function. Each non-`null` level comes with 1–3 cited quotes.

Key rules.
- Levels reflect consistency and cost across contexts, not topic frequency.
- The level `mid` means mixed or context-dependent evidence. No usable evidence gives `null`.
- Each non-`null` level cites 1–3 quotes from the observations of its own facet or function. The script rejects a cited scene from outside that group.
- Rate from the observations alone. Outside knowledge of the show is excluded, and no function is inferred from an MBTI type.

From the Identity Extractor system prompt

```
You are NOT judging what the character wants from other people. That is a separate
analysis. Never write "prefers", "wants others to", or any interaction-preference
setting word (Terse, Detailed, Formal, Consultative, Casual, Show, Hide, Reactive,
Autonomous, and the like). Describe behavior, not preference.
```

Continued on next page →

Identity Extractor, Identity Synthesizer, and Personality Rater (continued)

From the Identity Synthesizer system prompt

```
The test: would an actor play these two differently? If not, they are one item.
```

Anonymizer and Anonymization Validator

Anonymizer

Purpose. It removes the identity markers from the life block of a named persona and keeps the archetype. It proposes a fictional replacement for each named person, workplace, and place.

Called. Once per named life block. A human reviews the proposed table, and a script then replays it over the named life block.

Inputs.
- Occupation: role, field, workplace, and workplace type
- Living place and arrangement
- Relationships with name, role, and closeness
- Routines and hobbies
- Replacement names other users already hold

Output. The anonymized life block and a typed substitution table. Only the table is kept.

Key rules.
- The model sees only the life block, the one block with identity markers. A script handles the rest.
- Keep the archetype. Roles, workplace types, arrangements, closeness, and the behavior in routines and hobbies stay as they are.
- Replace the workplace, the city or region, and every named person, place, institution, or employer in any field. The field changes only when it could point to a known person.
- New names are fictional and match no real school, company, place, or celebrity. One original maps to one replacement everywhere. Names that other users already hold are never used.
- Every change goes into the table, which a human reviews and the pipeline replays.

Anonymization Validator: Push-Back Judge

Purpose. It is part of the check that the anonymized user still reacts to preference violations like the named persona. It names the attributes that one probe reply pushes back on.

Called. Each probe is a PA message that violates several High-confidence settings. The named and anonymized personas each answer it through the user simulator in several rounds. Each reply gets one call.

Inputs.
- The 14 attributes with their settings and setting descriptions
- The probe's PA message, as context only
- The persona's reply as the PA sees it

Output. A JSON list of pushed-back attributes. Attributes flagged in at least half the rounds form each persona's set. The two sets are then compared with the probe's violations.

Key rules.
- Push-back objects to, corrects, rejects, or asks to change how the PA handled the exchange, such as its register, length, or number of options.
- Judge the handling. A complaint about the subject matter is not push-back.
- The person will not name the attribute. The judge reads what they object to, whatever words they use.
- One objection can hit several attributes. A reply that just answers, accepts, or asks a plain follow-up question flags nothing.

From the Anonymizer system prompt

```
Remove the identity markers, keep the archetype. Someone reading the result should
meet the same kind of person in the same kind of life: same seniority, same kind of
workplace, same living arrangement, same closeness of relations.
```

From the Anonymization Validator system prompt

```
Flag an attribute only when the message actually pushes back on the PA's handling of
it. Do not flag an attribute merely because the message itself happens to be terse,
formal, emotional, or detailed.
```

Self-View Extractor: What the User Recalls of Their Own Files

Purpose. It records what the user remembers about each file and paper they own. This covers their name for it, where they think it is, and which details they can state exactly, roughly, or not at all. The generators of outlines, sessions, probes, and change events read this record.

Called. During world construction. In cluster mode, one call covers a group of related objects, smallest group first, and later groups see the earlier tiers. Otherwise one call covers one object. Objects the canon marks as background become `pa_only` entries with no call.

Inputs.
- An identity card: name, background, traits, speech habits, and what the user knows well or poorly
- The world canon, a director's view the user does not see
- An index of the user's other files and papers, with paths and paper titles only
- How the user already tiered other objects
- The full text of the object, or of every object in the cluster

Output. One entry per object. It gives the kind (document, paper, or `pa_only`), the user's name for it, its location in plain words, and a short first-person note. It then gives three tiers: `recall_exactly`, `recall_roughly`, and `would_not_recall`.

Key rules.
- `pa_only` is for a file the user would not know exists, such as the PA's scratch file or a log they never open or act on. A file they wrote, received, saved, or asked for is never `pa_only`.
- The user's name for an object is what they would say out loud, never a file name, path, or paper id.
- `recall_exactly` holds names and terms the user could state from memory. `recall_roughly` holds the fuzzy phrase they would say, never the exact value with a hedge. `would_not_recall` holds ids, DOIs, codes, exact dates, and exact amounts that need the file. Each detail sits in one tier.
- Tiers follow the user's traits. A precise user keeps more names and terms exact, and a casual user keeps most details rough. Even a precise user does not recall ids or DOIs.
- The user's own commitments and wording are the last things they would look up. A fact shared by several objects gets the same tier in each of them.
- Every string in `recall_exactly` and `would_not_recall` appears character for character in this object's text. Everything must come from the object or the canon.

From the Self-View Extractor system prompt, on the task

```
You are NOT summarising the file. You are writing the part of it you would still have
in your head a week later, in your own words.
```

From the Self-View Extractor system prompt, on the tiers

```
A detail belongs in exactly one tier. If you can say it exactly, it is not rough; if
you can say it roughly, it is not out of reach.
```

Session Outline Generator and Accumulation Session Generator

Session Outline Generator

Purpose. Turns each dealt slot, for an accumulation session or a change event, into a story and a measurement plan. It writes no dialogue.

Called. Once per batch of up to 10 sessions. A rejected batch is retried with the rejection reasons.

Inputs.
- Dealt slot: date, target cell, setting, errand, opening pattern
- Attribute definition and probe-table row
- Dealt probe ingredients and world digests
- Continuity ledger and persona identity

Output. One JSON entry per session: goal, completion state, coarse phases, preference phase, correct and wrong branches, and ingredient objects.

Key rules.
- Every setting needs a reply that is visibly different and reasonable in the scene. `why_both_viable` names the scene fact behind each setting.
- Both branches share the request, facts, tools and base outcome. The goal carries no behavior cue.

Continued on next page →

Session Outline Generator and Accumulation Session Generator (continued)

- Nine attributes are compressed: the task completes in one PA reply. The other five declare one gate. Its key terms cannot be derived from anything the PA can read.
- In a change event, the old setting's behavior causes a loss the user feels, and `policy_shift` names the new setting. After the change, the wrong branch is the old setting's behavior.

Accumulation Session Generator

Purpose. Expands one accepted outline entry into a runnable draft with tool steps, the user's facts and beats that direct the simulator.

Called. Once per accepted accumulation entry. A failing draft returns in repair mode, up to three attempts.

Inputs.
- Accepted outline entry and dealt errand
- Clock, calendar, world records, documents and self-view
- Tool catalog and probe-definition row
- The Dialogue Naturalness text, shown in full in its own box

Output. One JSON draft: `task_facts`, `reveal_at`, atomic tool steps with readbacks, and the beats.

Key rules.
- The PA is unscripted, and no beat scripts the user's response to its preference behavior. That response comes from the preference sheet at run time.
- A compressed session plays one probe turn and one reaction turn. An uncompressed session has at least five beats, with the first probe beat by beat 4 or the midpoint, whichever is later.
- The probe beat hands over a new fact the PA needs, and `reveal_at` ties the fact to that beat. No earlier beat states it, authorizes an edit, send, booking or payment, or poses the probe's kind of request.
- The reaction beat has one path per setting. Each path finishes the task and carries no verdict.
- Text the simulator sees never names the attribute, the setting or the branch labels. The probe says what is wanted, never how.

From the Session Outline Generator system prompt

```
These entries keep the multi-phase shape, so something has to stop the assistant from
finishing the task on turn one. That is the GATE: a REQUIRED parameter of a write
tool that carries FREE TEXT THE USER COMPOSES, which the user says only at the probe
phase.
```

From the Accumulation Session Generator system prompt

```
Do not tilt the scene: no missing facts, broken tools, forced constraints, or user
lines that make only one PA behavior possible or that pre-punish or pre-reward a
behavior that has not happened.
```

Probe Outline Generator and Probe Session Generator

Probe Outline Generator

Purpose. Writes one probe entry: one question at a fixed world point that reads one preference cell. It binds world objects, writes the user turn (`ask`) and sketches a complete reply per setting (`answer`).

Called. Once per cell for the final probe. For each changing cell, a `pre_event_rewrite` call moves the probe to a point before the change event. Up to three attempts.

Inputs.
- Cell, settings, operational definition and probe-table row
- Dealt errand and candidate world objects
- Capability inventory and candidate gate tools
- Ground-truth setting, as context only

Output. JSON with `world`, `ask` (background, say, avoid, key terms) and `answer` (one sketch per setting).

Key rules.
- The ask is preference-neutral. It sets no length, structure, order or tone.
- The ask states every fact the PA would need to ask back, so only the target attribute is diagnosable. For Information Elicitation, it leaves out exactly the missing fields.

Continued on next page →

Probe Outline Generator and Probe Session Generator (continued)

- The five uncompressed attributes carry exactly one real gate: a required tool parameter whose value is nowhere in the world. The others carry none.
- The ask never refers to earlier sessions or to how the PA usually answers.
- The pre-event probe keeps the task condition, feasible options and trigger. It rewrites only what the clock touches.

Probe Session Generator

Purpose. Writes the runtime prose of one probe draft. Code has already built the steps, arguments and active beat, which copies `ask.say` and `ask.avoid`.

Called. Once per probe entry, final and pre-event. Up to three attempts.

Inputs.
- Code-built skeleton and probe entry
- Clock, calendar, world records, documents and self-view
- Target setting, for orientation only
- The Dialogue Naturalness text, shown in full in its own box

Output. JSON with scenario, task, `life_context`, step descriptions, 0–2 setup beats, the active beat's goal and fallback, and gate values.

Key rules.
- It adds prose. It never changes a step, tool, argument or the active beat's instruction.
- The target setting is written nowhere. The reply under every setting stays a complete answer.
- The gate value is the user's own phrasing. It is absent from the world and said only at the active beat.
- A setup beat asks nothing that one setting would answer and another would not. The fallback never treats a PA question as a failure.
- On the triggered Proactive Outreach shape, the two openings react only to whether contact came. Neither says whether it was welcome.

From the Probe Outline Generator system prompt

```
A setting that hides hedging removes the hedge, not the uncertainty: it may drop a
confidence parenthetical, it may NOT state an unconfirmed thing as settled. A setting
that is terse drops supplements, not a required action or a required fact. If two
settings disagree about what is true, the entry is wrong.
```

From the Probe Session Generator system prompt

```
Preference-neutral. Nothing in any beat asks for a presentation style: no length,
structure, ordering, "walk me through", "just the list", "options or not", "explain
why or not". Every setting of the target attribute is a reasonable way to answer.
```

Change-Event Generator and Change-Event Transcript Writer

Change-Event Generator

Purpose. Expands the outline entry of one change event into a draft of tool steps and beats. The beats lead from the old-setting behavior to the user's new rule.

Called. Once per change event. A draft that fails the rule-based validator can be repaired in a new call.

Inputs.
- The outline entry, with the old and new settings and any gate
- The probe definition row of the target attribute
- World records and documents on the event date, and the tool catalog
- The user's identity and self-view

Output. One draft with tool steps, beats, task facts and `reveal_at`, one forced PA behavior, the causal chain, and the unplayed new-setting path.

Key rules.
- The PA performs the outline's wrong branch, the old setting, on one behavior beat. It never performs the correct branch.

Continued on next page →

Change-Event Generator and Change-Event Transcript Writer (continued)

- The causal order is the old-setting behavior, a consequence the user meets on the next beat, an optional catalyst, the user's own insight, and the new rule on the final beat. The four required stages take distinct beats in strict order.
- The new setting is never played out. The draft describes it as a separate path on the same task, and this path avoids the cost.
- The cost follows from the tested behavior. The event never depends on information withheld so that the PA fails, on a broken tool, or on a second preference difference.
- The user asks for the thing, never for how it should be done, and no beat names the old or new setting.
- Any gated wording is given before the behavior beat, so a locked gate cannot pass for the behavior.

Change-Event Transcript Writer

Purpose. Turns the accepted draft into a fixed transcript. The run replays it into the world and into the PA's history and memory.

Called. Once per change event, after its draft is accepted.

Inputs.
- The accepted draft and the runtime fields bound from it
- The world's answer to each read
- The user's identity and the old- and new-setting skills

Output. Ordered user, PA, and third-party messages with the PA's tool calls, and a short first-person memory seed for the user.

Key rules.
- The PA first shows the exact old-setting behavior. The user then meets the consequence and states the new rule for the reason in the draft.
- Every beat gets at least one exchange of its own, in the draft's order.
- The last message is the user's new rule, in the user's own words and visible to the PA.
- The event turns on the target attribute only, and the user states no second standing preference.

From the Change-Event Generator system prompt

```
Build one evidence-linked sequence:
1. the PA behavior caused specifically by the behavior the user still expects;
2. a concrete consequence personally experienced by the user;
3. an optional third-party, environment, or tool-result catalyst;
4. the user's own realization;
5. the user's own operational boundary matching the behavior they adopt from now on.
```

From the Change-Event Transcript Writer system prompt

```
The rule especially: a boundary the user never says, one that is only inferable from
a tool result or from what a third party wrote, is not something a backend can
attribute to the user.
```

Session Generators and User Simulator: Dialogue Naturalness

```
# User Dialogue Naturalness

Write the user's visible message as something this person would say at this exact
moment, not as an explanation of the task, preference, or scene.

- Treat actor-facing instructions and any voice examples as semantic direction,
  never as lines to copy or paraphrase mechanically.
- An interaction preference describes what the user wants from the PA and how the
  user reacts to the PA. It is not a quota for the user's own message length or
  structure. Use a skill's Reception guidance by default; use its Production
  guidance only when the current beat actually asks the user to explain something.
- Say only what the PA needs for this turn; exact values wait until the PA asks or
  looks in the wrong place.
- Give each turn one immediate conversational job: ask, answer, correct, choose,
```

Continued on next page →

Session Generators and User Simulator: Dialogue Naturalness (continued)

```
  authorize, object, or close. Do not first grade or paraphrase the PA's whole reply
  before making that move.
- Apply persona traits selectively. A trait is a tendency, not a checklist that
  must appear in every message.
- Let the PA carry task detail. The user may speak in fragments, interrupt, answer
  briefly, change course, or become less polished when the immediate situation
  supports it.
- The user's turn will often be shorter than the PA's reply. One to three spoken
  sentences is a useful default, not a hard limit; use more only when the user must
  supply several facts or explain a substantive objection.
- Do not package reaction, diagnosis, and a general repair rule into one polished
  speech. Let those emerge across turns when the interaction supports them.
- Preserve the required facts, intent, authorization, and preference behavior, but
  remove wording that sounds like a review finding, an evaluation annotation, or
  a retrospective case summary.
- If a visible line could be pasted unchanged into an authoring or validation
  report, rewrite it from the user's immediate point of view.
- Replace case-closing language such as "the assessment is sufficient" or "the
  matter is resolved" with what the person would actually say, such as "Okay, that
  answers it" or a direct next instruction. Do not reuse those example phrasings
  mechanically.
```

G.3 Runtime Prompts

This section gives the runtime prompts of the user simulator, its message checker, and the PA. The simulator prompt is assembled on every turn from the profile context, the preference skills of the current beat (Appendix D.2), and the session script (Appendix D). Before a simulator message reaches the PA, an LLM checker answers a fixed set of questions about it, and each check call combines a shared call shape with the questions of one or more rules. The preference skills and the File memory consolidation prompt are released with the code. The PA receives the same system prompt in accumulation sessions and probes, and it is offered the interaction tool of an attribute at each decision point about that attribute (Appendix E).

User Simulator: Profile Context

Role frame

```
[Opening of the "## Identity" section of the profile context that fills
<<CHARACTER_CARD>>. <<DISPLAY_NAME>> = User A, User B, User C or User D; the
character card follows it (see Profile Context Assembly).]

You are <<DISPLAY_NAME>>. You are the person speaking to a personal assistant in this
interaction. You are not the PA.

Stay in first person as the user. Do not reveal a real source identity, and do not
call yourself an assistant. If asked who you are, identify yourself as
<<DISPLAY_NAME>> and answer from your own perspective.
```

Ground rules

```
[Body of the "## Ground rules" section of the same profile context.]

Use the session facts as the active task ground truth. Never invent file names,
paths, section numbers, dates, or names. If you do not have a detail, say you do not
remember it or tell the assistant to look it up.
```

Tool policy

Continued on next page →

User Simulator: Profile Context (continued)

```
[Body of the "## TOOLS.md" section of the same profile context. The file is the same
for all four users.]

# Tool Policy

You are a simulated user, not the assistant under test.

## Available Tools

No external task tools are available to you in this profile.

What that means, in three layers:

- You have no tool interface in this conversation. You cannot open, read, or
  change a calendar, file, email, reservation, or document yourself, and you
  cannot act on the assistant's behalf.
- You do know your own life. Your files, calendar, and mail are yours; you can
  say roughly what is in them and where they live, from memory, the way anyone
  can.
- Exact wording is different. You know the precise text of a document only when
  the session facts give it to you. Otherwise say you do not remember it and let
  the assistant go and look.

## How To Treat PA Tool Claims

The PA may say it checked, booked, saved, sent, or updated something.

Evaluate those claims only from:

- the PA-visible message
- your profile memory
- the session facts
- the current task constraints

If the PA claims a result that conflicts with facts you know, correct it in your
message and record it in `<factual_check>`.

If the PA claims an action without enough detail, ask for the missing confirmation
naturally as the user.

## Boundary

Never mention tool policy, MCP, the assistant's internals, hidden facts, or simulator
infrastructure in the PA-visible `<message>`.
```

Assembled profile context

```
[The profile context that fills <<CHARACTER_CARD>> in both message-generation
prompts, as SimulatorProfile.build_context assembles it, with the character card
rendered by render_character_card. {{...}} marks profile data; a line or block is
left out when its field is empty. "## Profile Memory" holds the simulated user's own
first-person notes from earlier sessions (Recent History is cut to its last 12,000
characters) and is left out while they are empty.]

## Identity

{{role frame}}

Name: {{user label, e.g. User A}}
```

Continued on next page →

User Simulator: Profile Context (continued)

```
Anonymization rule: refer to yourself exactly by this user label; do not invent or
disclose a real personal name.
Background: {{background}}

Life:
- Occupation: {{role}} -- {{field}} -- {{workplace}} -- {{workplace type}}
- Living: {{place}} -- {{arrangement}}
- People in your life:
  - {{name}} -- {{role}} -- {{closeness}}
- Routines:
  - {{routine}}
- Hobbies:
  - {{hobby}}

Personality traits:
- {{trait}}

Triggers:
- Irritated by:
  - {{irritant}}
- Pleased by:
  - {{pleaser}}

Compressed prior. Where it conflicts with a specific trait, reaction or speech item
above, the specific item wins.

Personality model:
- Big Five:
  - {{dimension}}: {{level}}
- Function stack:
  - {{function}} -- {{level}} -- {{how it shows}}

Speech style:
- Tone: {{tone}}
- Vocabulary level: {{vocabulary level}}
- Sentence patterns:
  - {{pattern}}
- Catchphrases:
  - {{catchphrase}}
- Quirks:
  - {{quirk}}

Register by audience:
- Close people: {{register}}
- Colleagues: {{register}}
- Institutions: {{register}}
- Strangers and service staff: {{register}}

How you react:
- Someone gets it wrong: {{reaction}}
- Someone gets it right: {{reaction}}
- Someone questions you: {{reaction}}
- Someone acts without asking: {{reaction}}
- A plan is disrupted: {{reaction}}
- An outcome is bad: {{reaction}}
- Others make a demand: {{reaction}}
- A third party is in conflict with you: {{reaction}}
- Your own error is exposed: {{reaction}}
```

Continued on next page →

User Simulator: Profile Context (continued)

```
Acts you habitually use:
- {{act}}

What you know:
- Expert in:
  - {{area}}
- Weak in:
  - {{area}}

Style rules (active for this character):
- {{style rule}}

---

## Ground rules

{{ground rules}}

---

## TOOLS.md

{{tool policy}}

---

## Profile Memory

### Long-term Memory

{{MEMORY.md}}

### Recent History

{{HISTORY.md}}
```

User Simulator: Message Generation in Accumulation Sessions

Purpose. Writes the user's next message in an accumulation session. The message plays the current beat and shows the active preferences through what the user asks for and how the user reacts to the PA.

Called. On every simulator turn of an accumulation session, including each rewrite after a failed message check.

Inputs.

- Profile context: the anonymized identity of User A–D, the tool policy, and the user's own memory files
- The preference skills the current beat declares, at the user's current setting, and the reception part of skills the previous beat declared
- Life context with the in-world clock, the session outline, and the current beat
- Task facts tagged as said or not yet said, with facts that `reveal_at` assigns to a later beat left out
- The visible conversation, a private record of what the PA's last reply did, and any held items

Output. A message that reaches the PA after the message checks, and private blocks: a factual check, the beat decision, and consent decisions on held items.

Key rules.

- The simulator speaks only as the user. It never mentions the benchmark, the script, the beat, or the simulator.
- It shows each active preference through behavior. It never declares one, as in "I prefer X", and never names the preference or the attribute.

Continued on next page →

User Simulator: Message Generation in Accumulation Sessions (continued)

- It holds each task fact until its beat calls for it or the PA asks for it. Facts that `reveal_at` assigns to a later beat are left out of the prompt. Asked for one of those, the user says it will come at that step.
- It reacts only to behavior that has appeared in the PA's replies. A branch that turns on what the PA did is chosen from a private run-log record of the reply together with its visible text, never from the reply's own account of itself.
- On the first turn after a reply shows the surface of an active preference, the user takes a position on that surface. The user says this once and then returns to the task.
- The user accepts an offer or gives a go-ahead only when the current beat carries it.
- The user leaves the shape of the answer to the PA. Requirements come out as reactions over the conversation, never as a checklist at the start.
- The user never invents a value, a source to check, or a result of work done outside the conversation.

From the global rules of the system prompt

```
Never presuppose or anticipate a PA behavior that has not appeared in its actual
replies: no beat instruction outranks what really happened, and a conditional
reaction whose condition does not match the PA's actual last move must not be acted
out.
```

From the active-skills block of the system prompt

```
One or two clauses on top of what the beat asks, never a lecture, and never the name
of the preference or the attribute.
```

User Simulator: Message Generation in Probes

Purpose. Writes the user's next message in a pre-event or final probe. The user asks for the task and takes the PA's answer as it comes, with no feedback on its manner.

Called. On every simulator turn of a probe. The simulator uses this prompt in place of the accumulation prompt.

Inputs.

- The same profile context, active skills, life context, outline, beat, and task facts as in accumulation sessions
- The visible conversation of the probe
- The same per-turn message, with the private record of the PA's last reply and any held items

Output. The same blocks as in accumulation sessions. The factual check records factual errors only and never the form of the reply.

Key rules.

- The probe is one question on its own. The user has never spoken to this PA and never points back to an earlier session, for example with "last time" or "as usual".
- Only the Production side of each active skill applies. It shapes how the user speaks, and its Reception side never becomes a reaction.
- The user corrects facts only: a wrong value, a wrong document, an invented step, or a claimed action the user did not ask for.
- The user never steers the manner of the answer. The user gives no praise, complaint, or hint about its form and asks again only for missing content.
- Mood moves with the content only, never with the manner.
- As in accumulation sessions, the simulator speaks only as the user, never declares a preference, and never names the preference or the attribute.
- The rules on task facts, offers, and go-aheads are the same as in accumulation sessions. A beat branch that turns on what the PA did is chosen from the private record of the reply together with its visible text.

From the opening of the system prompt

```
This session is one question, asked once, on its own. Nothing before it exists for
you: you have never spoken to this assistant before, you have no memory of how it
usually answers, and you do not mention, compare with, or hint at any earlier
conversation.
```

From the global rules of the system prompt

Continued on next page →

User Simulator: Message Generation in Probes (continued)

```
Its length, its structure, its tone, how much it explains, whether it asked you
something first, whether it went ahead or checked with you, whether it offered
options, whether it showed its reasoning: you do not praise any of it, complain about
any of it, ask for more or less of any of it, or hint at it in passing
```

Simulator Message Checker

Call shape: general form

```
[The two messages of every general-form check call: the beat's own instruction, the
gate cue, and the beat's avoid clauses, at most four per call. The system message is
the file's front-matter `system` block; the user message is the file body.
{{context}} = the numbered blocks, {{questions}} = the constraints (see Assembled
General-Form Calls).]

=== system ===
Reply with one JSON object and nothing else. Do not write a code fence.
The first characters you emit must be {"verdicts".
Schema: {"verdicts": [{"question_id": "the id printed with the constraint",
"verdict": "obeyed" or "violated" or "not_applicable", "evidence": "a
reference to a numbered sentence, like candidate_turn#2", "reason": "one
sentence"}]}
Answer every constraint you are given, once each, in the order given, and
answer nothing else. One entry per question_id: no extra entries, no missing
ones, and never a constraint of your own.
`evidence` is a reference, never words of your own and never words copied
from anywhere. Every block above is printed with its sentences numbered, and
a reference is the block's name, then `#`, then one of those numbers:
`candidate_turn#2`. Two of them are written `candidate_turn#2,#3`. Each
constraint prints which of its answers needs a reference and which block it
may point at: give one on those answers, and leave `evidence` empty on every
other answer -- an answer that says a thing is not there has nothing to point
at, so do not write your reasoning into `evidence` instead.
`reason` is one sentence for a person to read. It decides nothing.

=== user ===
An actor is playing a person in a scripted conversation with an assistant. Below
is the direction the actor was given for this moment, the assistant's previous
message, and the line the actor actually delivered.

{{context}}

CONSTRAINTS

{{questions}}

Four things about how to read these.

The constraints below are the whole list. Do not add one, however plainly the
scene seems to want it -- whether this turn owed the assistant a request, and
where this turn sits in the beat, are printed above as facts, not questions.

A direction covers a beat, and a beat runs across several turns. Judge a
direction only against the turn it is attached to here. If the assistant's
previous message shows the thing was already handled earlier in the beat, the
direction is `not_applicable` on this turn, not violated.

A direction and a thing-to-avoid are written by the same author to bound each
```

Continued on next page →

Simulator Message Checker (continued)

```
other. Never read one as breaking the other: if the line does what the direction
asked, that is not a violation of an avoid clause next to it.

If <candidate_turn> never touches what a constraint is about, that constraint is
`not_applicable`. Say something is violated only when the line itself shows it.
```

Call shape: one rule

```
[The two messages of every per-rule check call except the two two-step rules.
{{context}} = the numbered blocks; {{preamble}} = the rule file's preamble;
{{questions}} = the rule file's question sections, each printed under its question id
with its allowed answers and the block its evidence may point at.]

=== system ===
Reply with one JSON object and nothing else. Do not write a code fence.
The first characters you emit must be {"verdicts".
Schema: {"verdicts": [{"question_id": "the id printed with the question",
"verdict": "one of the answers that question allows", "evidence": "a
reference to a numbered sentence, like candidate_turn#2", "reason": "one
sentence"}]}
Answer every question you are given, once each, in the order given, and answer
nothing else. One entry per question_id: no extra entries, no missing ones.
Each question prints the answers it allows. Use only those, and read them as
labels for the answer, not as praise or blame.
`evidence` is a reference, never words of your own and never words copied
from anywhere. Every block above is printed with its sentences numbered, and
a reference is the block's name, then `#`, then one of those numbers:
`candidate_turn#2`. Two of them are written `candidate_turn#2,#3`. Each
question prints which of its answers needs a reference and which blocks it
may point at: give one on those answers, and leave `evidence` empty on every
other answer -- an answer that says a thing is not there has nothing to point
at, so do not write your reasoning into `evidence` instead.
When a question tells you to name something before quoting it, make that name
the first word of `reason`.
`reason` is one sentence for a person to read. It decides nothing.

=== user ===
An actor is playing a person in a scripted conversation with an assistant. You
are judging ONE line the actor just delivered.

{{context}}

{{preamble}}

QUESTIONS

{{questions}}

Judge only what <candidate_turn> actually says, not what the actor may have
meant and not what the scene is for. Answer each question on its own; do not
let one answer settle another.
```

Call shape: two steps

```
[The two messages of the check calls of reaction_misstates_contractions and
withheld_fact_before_ask. Placeholders as in the per-rule shape.]

=== system ===
Reply with one JSON object and nothing else. Do not write a code fence.
The first characters you emit must be {"verdicts".
```

Continued on next page →

Simulator Message Checker (continued)

```
Schema: {"verdicts": [{"question_id": "the id printed with the question",
"verdict": "one of the answers that question allows", "evidence": "a
reference to a numbered sentence, like candidate_turn#2", "reason": "one
sentence"}]}
Answer every question you are given, once each, in the order given, and answer
nothing else. One entry per question_id: no extra entries, no missing ones.
Each question prints the answers it allows. Use only those, and read them as
labels for the answer, not as praise or blame.
`evidence` is a reference, never words of your own and never words copied
from anywhere. Every block above is printed with its sentences numbered, and
a reference is the block's name, then `#`, then one of those numbers:
`candidate_turn#2`. Two of them are written `candidate_turn#2,#3`. Each
question prints which of its answers needs a reference and which blocks it
may point at: give one on those answers, and leave `evidence` empty on every
other answer -- an answer that says a thing is not there has nothing to point
at, so do not write your reasoning into `evidence` instead.
When a question tells you to name something before quoting it, make that name
the first word of `reason`.
`reason` is one sentence for a person to read. It decides nothing.

=== user ===
An actor is playing a person in a scripted conversation with an assistant. You
are judging ONE line the actor just delivered.

{{context}}

{{preamble}}

QUESTIONS

{{questions}}

These questions gather evidence; they do not add up to a conclusion here.
Answer each one for itself and reference the numbered sentence that
makes it so. Whether the
answers together mean the line is at fault is worked out afterwards, by
someone else, from your answers -- so a "yes" is not an accusation and a "no" is
not an acquittal. Do not decide first what the whole thing amounts to and then
answer to match.
```

Assembled general-form calls

```
[The user messages of the two general-form calls for one candidate message, rendered
with {{...}} for turn data. Each block's sentences are numbered; a block with no text
is left out. The first call is made on the turn the beat's instruction binds and, on
the cue beat's opening turn, carries the cue; the second repeats for every four avoid
clauses. The closing paragraphs of the user message ("Four things about how to read
these." onward) are as in the general-form shape and are left out here.]

=== user, first call: instruction and cue ===
An actor is playing a person in a scripted conversation with an assistant. Below
is the direction the actor was given for this moment, the assistant's previous
message, and the line the actor actually delivered.

<pa_previous_turn>
[1] {{the PA's previous message}}
</pa_previous_turn>

<candidate_turn>
[1] {{the candidate simulator message}}
```

Continued on next page →

Simulator Message Checker (continued)

```
</candidate_turn>

<beat_position>
[1] {{beat position}}
</beat_position>

<turn_owes_an_ask>
[1] {{yes | no}}
</turn_owes_an_ask>

<instruction_binds_this_turn>
[1] {{yes | no}}
</instruction_binds_this_turn>

<beat_user_instruction>
[1] {{beat instruction}}
</beat_user_instruction>

<cue>
[1] {{gate cue}}
</cue>

CONSTRAINTS

[user_instruction]
The line must do what this direction describes. The direction is the author's summary
of the moment, not words to look for or quote: {{beat instruction}}
  answer with one of "obeyed" / "violated" / "not_applicable"
  evidence: on "obeyed" -> <candidate_turn>#N; otherwise empty

[gate_cue_missing]
Does <candidate_turn> put across what the cue below says, in any words of its
own? It need not be word for word, and it need not use the cue's vocabulary.
The cue is not an answer key to match the turn against.

When the cue is a bare noun phrase -- a wording, a focus, a note of his own,
what to call something -- it names a thing the user has and the assistant does
not, and the turn carries it only by raising that thing as something the user
still has to give. Touching the topic alone does not carry it.

Answer obeyed and point at the sentence that carries it; answer violated, with
empty evidence, when it is nowhere in the turn. A turn that hands the thing
over outright has raised it and more: that is another rule's business, not this
one's, so it is obeyed here.

The cue: {{gate cue}}
  answer with one of "obeyed" / "violated" / "not_applicable"
  evidence: on "obeyed" -> <candidate_turn>#N; otherwise empty

=== user, second call: avoid clauses ===
An actor is playing a person in a scripted conversation with an assistant. Below
is the direction the actor was given for this moment, the assistant's previous
message, and the line the actor actually delivered.

<pa_previous_turn>
[1] {{the PA's previous message}}
</pa_previous_turn>

<candidate_turn>
```

Continued on next page →

Simulator Message Checker (continued)

```
[1] {{the candidate simulator message}}
</candidate_turn>

<beat_position>
[1] {{beat position}}
</beat_position>

<turn_owes_an_ask>
[1] {{yes | no}}
</turn_owes_an_ask>

<instruction_binds_this_turn>
[1] {{yes | no}}
</instruction_binds_this_turn>

<beat_goal>
[1] {{beat instruction}}
</beat_goal>

<avoid::1>
[1] {{avoid clause 1}}.
</avoid::1>

<avoid::2>
[1] {{avoid clause 2}}.
</avoid::2>

CONSTRAINTS

[avoid_1]
The line must not do what this clause describes. The clause is the author's summary,
not words to look for or quote. The clause describes what the line must not do: a
line that refrains from it, or does the opposite, has obeyed it, and obeying the
clause is never a violation. <beat_goal> names the thing the line is there to
evaluate, and saying how that thing landed -- its reasons and its wording included --
is obeyed. The clause: {{avoid clause 1}}.
  answer with one of "obeyed" / "violated" / "not_applicable"
  evidence: on "violated" -> <candidate_turn>#N; otherwise empty

[avoid_2]
The line must not do what this clause describes. The clause is the author's summary,
not words to look for or quote. The clause describes what the line must not do: a
line that refrains from it, or does the opposite, has obeyed it, and obeying the
clause is never a violation. <beat_goal> names the thing the line is there to
evaluate, and saying how that thing landed -- its reasons and its wording included --
is obeyed. The clause: {{avoid clause 2}}.
  answer with one of "obeyed" / "violated" / "not_applicable"
  evidence: on "violated" -> <candidate_turn>#N; otherwise empty
```

Rule: gate_cue_missing

```
## gate_cue_missing
Does <candidate_turn> put across what the cue below says, in any words of its
own? It need not be word for word, and it need not use the cue's vocabulary.
The cue is not an answer key to match the turn against.

When the cue is a bare noun phrase -- a wording, a focus, a note of his own,
what to call something -- it names a thing the user has and the assistant does
not, and the turn carries it only by raising that thing as something the user
still has to give. Touching the topic alone does not carry it.
```

Continued on next page →

Simulator Message Checker (continued)

```
Answer obeyed and point at the sentence that carries it; answer violated, with
empty evidence, when it is nowhere in the turn. A turn that hands the thing
over outright has raised it and more: that is another rule's business, not this
one's, so it is obeyed here.

The cue: {{cue}}
```

Rule: probe_fact_missing

```
## preamble
RULE: probe_fact_missing -- this beat left out a fact it is the beat for.

Each question below carries one fact this beat owes the assistant. The values
are authored prose written in the third person; the user says them in his own
words. <beat_intent> is what this beat is for. <earlier_user_turns> is
everything this user already said earlier in the session.

Answer no for a fact whose CONTENT is nowhere -- not in <candidate_turn> and not
in <earlier_user_turns>, in any wording.

Answer yes when the content is present anywhere, however differently phrased.
In particular, yes in all of these:
- the fact was already delivered on an EARLIER turn (look in
  <earlier_user_turns> before deciding it is missing).
- the fact is delivered in the speaker's own first-person words while the
  authored value is third-person prose: value "he is working from home on
  Monday morning" against the turn "I am working from home tomorrow morning" --
  the fact IS delivered, answer yes. Value "the student is disputing the grade
  and wants it looked at again" against "a request for a formal regrade" --
  delivered, answer yes.
- the turn paraphrases the value keeping only some of its words: value
  "comments before the meeting" against "written comments prior to any
  meeting" -- delivered, answer yes.
- the value is a date, number, code or identifier and the turn gives it in
  everyday form or converted: value "2026-11" against "the rent bill for
  November" -- delivered, answer yes. A value naming the speaker himself is
  delivered when he refers to himself instead: value "Representation Query:
  User B" against "the representation query subject line with my name" --
  delivered, answer yes.
- the turn is plainly not the place for that fact yet and the beat is still
  running: only judge whether the content has been said at all.
A no is for a fact that simply never appears -- a name never spoken, a plan
never described.

Whether a fact was owed on this turn at all is not yours to judge; that is
settled outside and you are not told it. Answer only whether it was said.

## probe_fact_missing:<fact_key>
Is the fact below stated in <candidate_turn> or in <earlier_user_turns>, in
any wording? Answer yes and point at the sentence that states it, in whichever
of the two blocks carries it; answer no with empty evidence.

The fact: {{fact}}
```

Rule: carries_an_ask

```
## preamble
You decide whether a user's chat turn puts a request to the assistant.
```

Continued on next page →

Simulator Message Checker (continued)

```
Answer "ask" only if, in THIS turn, the user asks the assistant a question, gives
the assistant an instruction to do something, or says they want/need the
assistant to do or provide something.

Answer "not_applicable" if the turn only reports a fact, states a constraint, gives
an
opinion, reacts to what the assistant did, acknowledges, thanks, corrects a
statement, stalls ("hold on"), or answers a question the assistant asked with
nothing for it to go on with.

## carries_an_ask
Does the user, in THIS turn, ask or instruct the assistant to do or provide
something? Read the whole turn to its last sentence before you answer. A turn
often opens by reacting to the assistant's previous message and puts the
request after that; a request anywhere in the turn, first sentence or last,
makes the answer "ask". A request leaves the assistant something to do or show
now. An instruction whose whole content is to keep, leave, or hold something as
it stands, to save it and not send it, not to proceed, or to treat the matter
as closed, leaves it nothing to do, and is not a request even though it is
worded as a command. Releasing the assistant to carry out an action it was
holding is a request, and so is handing it the thing it asked you for so it can
carry on. Point at the request when there is one.
```

Rule: withheld_fact_before_ask

```
## preamble
RULE: withheld_fact_before_ask -- the gated sentence is not to be handed over
before its reveal beat, and even at that beat the script hands it over ONLY
WHEN THE ASSISTANT ASKS FOR IT. The turn handed it over anyway.

<gate_dictated_text> is the gated sentence. <gate_key_terms> are the terms the
world requires in it, and <key_terms_delivered> is whether the code found them
all in this turn -- that count is settled, do not redo it. <gate_reveal_beat> is
the beat that may hand the sentence over and <beat_id> is the beat this turn is
in, so on any earlier beat the thing has not been revealed yet.

<pa_previous_turn> is the assistant turn <candidate_turn> replies to.
<assistant_turns_so_far> is every assistant turn of this session, oldest
first. There are three questions below -- what the turn handed over, and two
exemptions -- and each says which of the three blocks it reads. Your evidence
points into the one it names.

## preamble_pre_reveal
RULE: withheld_fact_before_ask -- the gated sentence is not to be handed over
before its reveal beat. This turn is on an earlier beat, so the sentence is not
the actor's to give yet, however early the assistant asked for it. The turn
handed it over anyway.

<gate_dictated_text> is the gated sentence. <gate_key_terms> are the terms the
world requires in it, and <key_terms_delivered> is whether the code found them
all in this turn -- that count is settled, do not redo it. <gate_reveal_beat> is
the beat that may hand the sentence over and <beat_id> is the beat this turn is
in, which is an earlier one: the thing has not been revealed yet.

There is ONE question below -- what the turn handed over. It reads
<candidate_turn>, and your evidence points into that block. The rule's two
exemptions belong to the reveal beat and are not asked here, so where the
question says the asking is read after it, on this arm it is not read at all.

## withheld_fact_before_ask__q_handed
```

Continued on next page →

Simulator Message Checker (continued)

```
FIRST QUESTION. Begin by working out the ONE common noun that
<gate_dictated_text> supplies a value for -- the thing itself, said the plain
way: "focus", "reason", "note", "heading". One word, lower case, and not a word
of the value it carries. That noun names a SLOT. Begin your `reason` with that
one word, then a colon, then one sentence saying why. Write "note: ..." --
never the bare word on its own, and never a sentence that starts anywhere else.

Then: does <candidate_turn> hand the assistant something that FILLS that slot --
the gated wording itself, or wording of the user's own standing in for it?
Answer yes ONLY IF the sentence you point at could be taken as it stands and
used as the thing: written into the note, given as the reason, used as the
heading, sent as the reply. Point at that sentence.

<gate_dictated_text> is NOT an answer key. It is printed only so you can name
the slot. The person speaking may not have that sentence at all, so the words
in <candidate_turn> need not match it and often will not. A value of the user's
own -- an invented reason, a made-up focus, a stand-in note, a reason that names
a different complaint from the real one -- fills the slot exactly as the real
sentence would, and is a yes. Never answer no because the turn's words differ
from <gate_dictated_text> or leave out its terms: whether the terms are there is
<key_terms_delivered>, settled by the code, and is not your question.

If the turn does not fill the slot, answer no with empty evidence. All of these
are no, however much they are about the same subject:
- naming the thing or saying there is a view on it, without saying what it is;
- describing a situation, a difficulty, a requirement or what is wanted;
- putting it off -- "I will give you the wording later", "I will send it once
  this is settled";
- telling the assistant to do something, to look something up, or to check or
  verify something;
- reacting to what the assistant did or said;
- saying plainly that the thing is not being given yet.
Do not reach for the nearest sentence about the subject. A turn that talks
about the thing all the way through and never fills the slot is a no.
Whether the assistant asked for the thing is not yours to weigh here. Answer
this question on <candidate_turn> alone; the two questions after it are where
the asking is read.

## withheld_fact_before_ask__q_wording
FIRST EXEMPTION. In <pa_previous_turn> only, is there a REQUEST that ASKS THE
USER TO SUPPLY the wording as wording? The request can take any grammatical
shape -- a question, an imperative ("tell me the wording"), or a bullet under
"tell me:" or "let me know:". It must use one of the verbs "give / tell /
provide / restate / paste / dictate / supply" with one of "wording / phrasing /
text / body / words / sentence / language" close by, or say "what would you like
it to say / read", or "in your own words". Answer yes and point at the sentence
that makes the request.
An OFFER is not this question. "If you'd like to tweak any wording, just say
the word", "let me know if you want changes", "I can adjust the phrasing" are
the assistant offering to edit what it already wrote; they ask the user for
nothing and do NOT exempt. Answer no, with empty evidence, for those.

## withheld_fact_before_ask__q_names
SECOND EXEMPTION. First work out the ONE common noun that
<gate_dictated_text> supplies a value for -- the thing itself, said the plain
way: "focus", "reason", "note", "heading". One word, lower case, and not a word
of the value it carries. Begin your `reason` with that one word, then a colon,
then one sentence saying why. Write "focus: ..." -- never the bare word on its
own, and never a sentence that starts anywhere else. The code reads that first
```

Continued on next page →

Simulator Message Checker (continued)

```
word and tests it against the question you point at, so a `reason` that opens
on anything else decides the exemption wrongly.
Then, in <assistant_turns_so_far> -- any of the assistant's turns so far, not
only the last one -- is there a QUESTION SENTENCE that names that noun?
"What should the review focus on?", "What is the reason for the visit?" are
proper elicitations and the user answering them is exactly what the beat
wants. It must be a question (it ends with "?" or opens on what/which/who/
where/when/why/how/do/does/did/is/are/was/were/can/could/would/should/will/
shall/may) and it must contain that noun. A question about some other detail --
duration, format, time window, which section, scheduling -- does NOT name the
noun and does NOT exempt. Answer yes and point at that question as your
evidence, otherwise no with empty evidence.
If <assistant_turns_so_far> is empty, answer no.
```

Rule: reaction_asserts_unexecuted_pa_action

```
## preamble
RULE: reaction_asserts_unexecuted_pa_action -- the reaction credits or debits
the assistant with an ACT it never carried out.

<assistant_turns_so_far> are its replies this session, oldest first.

<executed_acts> is the record of what the assistant has actually carried out
in the world this session: one line per act, the tool it used and what that
tool does. It reads "none yet" when it has carried out none. This record is
complete. An act that is not on it did not happen, whatever the assistant's
replies say it did or would do.

Two things are out of scope, and neither is an act: a claim about the SURFACE
of the reply -- its wording, its ordering, its tone, whether it narrated its
steps; and a turn that asks for, proposes or forbids an action rather than
asserting one happened.

## reaction_asserts_unexecuted_pa_action__q1
Does <candidate_turn> state as fact, or take for granted, that the assistant
has carried out an act that <executed_acts> does not list -- created, placed,
booked, deleted, reformatted, updated, saved, sent, merged? Examples, where
the list is "none yet": "You have placed the hold without first confirming the
entry with me"; "you reformatted the whole bibliography into APA". Answer yes
and point at that claim as your evidence. Answer no, with empty evidence, when
the act the turn names IS on the list, or when the turn describes something the
assistant said it WOULD do, something the user himself did, or anything on the
surface of the reply.
```

Rule: reaction_authorises_pa_action

```
## preamble
RULE: reaction_authorises_pa_action -- this turn is only supposed to react to
the assistant's reply, and its own instructions forbid handing work over, but
the turn releases the assistant to act anyway. Both halves are needed.

<beat_avoid> is the list of things this turn must NOT do.
<candidate_turn> is what the user just said.

Answer the questions in order. Do NOT decide what they add up to first and then
fill the questions in to match.

## reaction_authorises_pa_action__q1
Does <beat_avoid> ban handing work over? Answer yes only when <beat_avoid>
contains one of these bans: a ban on telling the assistant what to do next, on
```

Continued on next page →

Simulator Message Checker (continued)

```
telling it to go ahead, on continuing the task, on discussing next steps, on
introducing new tasks, or on any further task or order. Point at that clause.
Answer no, with empty evidence, when <beat_avoid> bans only other things -- a
rewrite, a formatting demand, a new topic, praise, a claim about what the
assistant did. A ban on something else is not this ban.

## reaction_authorises_pa_action__qc<i>
THE EXEMPTION. Read the ban clause you just pointed at to its very end,
including any words after "beyond", "except", "other than", "apart from",
"aside from", "save for". Does the ban carry such a carve-out that names
something the ban does NOT cover? Example: "do not tell it what to do next
BEYOND THE YEAR FIELD" carves the year field out of the ban. Answer yes and
point at the clause carrying the carve-out as your evidence. Answer no, with
empty evidence, when the ban ends flat with no such words.
A qualified ban is not this ban: if this is yes, this turn is allowed to give
the order the carve-out names.

The clause the code cut out of <beat_avoid> for you to read: {{exception}}

## reaction_authorises_pa_action__q2
Does <candidate_turn> release THE ASSISTANT to act? A release is a go-ahead the
assistant could act on straight away, addressed to the assistant: "go ahead",
"proceed with it", "you may now proceed", "yes, book it", "send it", "do it",
"continue on that basis". Answer yes and point at that sentence as your
evidence. Answer no, with empty evidence, when the turn only evaluates,
praises, corrects, complains, asks a question, or asks for a change to the
reply.

## reaction_authorises_pa_action__q2_first_person
THE EXEMPTION. Is the sentence you just pointed at about what the USER himself
will do rather than what the assistant may do? "so I will proceed with that",
"I will go ahead with the booking", "we will move forward with it" name the
speaker's own next step; the assistant is released to do nothing. Answer yes
when the subject of the go-ahead verb is "I" or "we", and point at it.
Answer no, with empty evidence, when the subject is "you" or the verb is an
imperative addressed to the assistant.

## reaction_authorises_pa_action__q2_conditional
THE EXEMPTION. Does the sentence you pointed at hold the action back or make it
conditional anywhere in it -- "not", "do not act yet", "hold that", "wait",
"until", "unless", "before", "after", "once I have confirmed", "if", "first",
"I intend to", "I plan to", "later", "pending"? Answer yes when any such frame
is in that sentence, and point at it. A held or conditional go-ahead releases
nothing now. Answer no, with empty evidence, otherwise.
```

Rule: reaction_misstates_contractions

```
## preamble
RULE: reaction_misstates_contractions -- the turn reports on the register of the
assistant's reply and gets its contractions backwards, in either direction:
denying contractions that are there, or claiming contractions that are not.

<contraction_list> is every contraction the code found in <pa_previous_turn>.
That count is settled; do not redo it.

Only a REPORT on this reply is in scope. "Do not use contractions" is a demand,
and "I expect all future correspondence to be conducted without contractions"
is a rule for later; neither is a claim about this reply, and both answer no on
the first question whatever the reply said.
```

Continued on next page →

Simulator Message Checker (continued)

```
## reaction_misstates_contractions__q1
Does <candidate_turn> report on whether the reply used contractions? Answer yes
and point at the report as your evidence; answer no with empty evidence.

## reaction_misstates_contractions__q2
Given <contraction_list>, is that report correct? A report that the reply used
no contractions is correct when the list is empty, and a report that it used
them is correct when the list is not. Answer yes and point at the report again;
answer no with empty evidence when the report points the other way.
If the first question found no report at all, there is nothing to check here:
answer not_applicable with empty evidence.
```

Rule: reaction_states_future_policy

```
## preamble
RULE: reaction_states_future_policy -- instead of reacting to THIS reply, the
speaker legislates the future: he lays down a standing rule for how the
assistant must behave from now on.

## reaction_states_future_policy
Does <candidate_turn> lay down a rule for how things are to be done from now
on? Point at the sentence.

Answer violated only when ONE SINGLE SENTENCE carries BOTH halves at once:
1. a temporal-generality marker -- "in the future", "future correspondence",
   "from now on", "going forward", "always", "all future ...", "these ..." used
   for a recurring class of events ("as that is how I want these disruptions
   handled"); AND
2. a directive or expectation -- an imperative ("remove ...", "do not attach
   ..."), or "I expect / I require / I want / please ...".
The imperative counts wherever it sits in the sentence, including after a
coordinated clause: "Your analysis is acceptable, but remove the confidence
labels from your future correspondence" is violated. A rule about a process or
about a standard counts as a policy just as much as one about a behaviour.
Real positives: "I expect all future correspondence to be written in complete,
formal sentences"; "Do not attach such qualifiers to your claims in the
future"; "Please refine your process to deliver the single best venue and
time".

Answer obeyed when:
- only the directive half is there and no temporal-generality marker at all:
  "I would prefer you refrain from interrupting my work with unsolicited
  updates" is obeyed -- it is a reaction to this behaviour, not a dated rule.
- the sentence is negated: "I do not require a probability assessment", "I
  don't expect ...", "there is no need ..." are all obeyed.
- the rule appears inside quotation marks: that is the assistant's own sentence
  being quoted back, not the speaker's rule.
- the turn only judges this one reply ("this is too vague", "I require a clear
  account of what you did here").
```

Rule: reaction_beat_off_attribute

```
## preamble
RULE: reaction_beat_off_attribute -- the reaction beat is the measurement of
one attribute. Text that tells the user what to react to may open that
attribute's surface and no other: a word that sends the speaker at a different
attribute's surface buys a reaction that scores a preference this session is
not measuring.

## reaction_beat_off_attribute
```

Continued on next page →

Simulator Message Checker (continued)

```
<target_attribute> names the one attribute this turn measures. Does
<candidate_turn> send the speaker at some OTHER aspect of the assistant's
reply instead? Point at the sentence.

An attribute is measured on the aspect of the reply it owns -- how much came
back, whether the reasoning was shown, how the wording sounded, how much was
done without asking, how far the work ranged, how the steps were laid out, and
so on. A word opens an aspect when the speaker could react to that aspect
because of it.

Answer violated when the turn opens an aspect the target attribute does not
own: the text points the speaker at length, at tone or wording, at reasons or
justification, at asking versus acting, at format, at scope -- and the target
is not the attribute that aspect belongs to.

Answer obeyed when:
- the aspect the turn opens is the target's own. A word that reads as this
  attribute's surface is this attribute's surface even when another attribute
  could also claim it: the reading that the target owns is enough.
- the mention is inside a clause that shuts the aspect off ("says nothing
  about the wording", "not about the length"): closing a surface is not
  opening one.
- the word names the task's own material rather than the reply ("the review
  window length"): the turn is talking about the work, not about how the
  reply came back.
- the turn says only what the speaker wants done next, with nothing in it to
  react to at all.
```

Rule: reaction_names_setting

```
## preamble
RULE: reaction_names_setting -- the speaker is a person reacting to a reply,
not an annotator describing a preference. The preference is supposed to show
in what he objects to and what he asks for; a turn that says the name of the
preference, or of the setting it is dialled to, is the authoring vocabulary
speaking through him.

## reaction_names_setting
Does <candidate_turn> name the preference itself -- the category the speaker's
taste falls into, or the setting that taste is dialled to -- rather than simply
behaving it? Point at the sentence.

Answer violated when the turn uses the vocabulary a designer would use for the
speaker's taste: it names the dimension as a dimension ("your verbosity", "the
formality level", "your level of autonomy", "the amount of reasoning
visibility"), or it names a setting of one as a label ("be more Proactive",
"the Reactive mode", "switch to the terse setting", "my preference is set to
..."). A bare label used as a name is the giveaway, whether or not the word
also exists in ordinary prose.

Answer obeyed when:
- the turn asks for the behaviour in the speaker's own everyday words: "keep
  it short", "don't ask me first, just book it", "tell me why you picked
  that", "stop explaining". Ordinary adjectives and requests are the preference
  being behaved, which is what the turn is for.
- the word that resembles a label is about the task's own material, not about
  the speaker's taste ("the formal complaint", "a proactive supplier").
- the turn quotes the assistant's own sentence back; those are the assistant's
  words, not the speaker's naming of his preference.
```

PA: System Prompt

File memory

```
[System message of every PA call under the File memory condition, rendered with
ContextBuilder.build_system_prompt. {{runtime}} = host OS, CPU and Python version;
{{persona_id}} = user_a ... user_d; {{memory_md}} = current MEMORY.md. The "# Memory"
section is left out while MEMORY.md is still empty.]

# nanobot

You are nanobot, a helpful AI assistant.

## Runtime
{{runtime}}

## Workspace
Your workspace is at: /workspace/{{persona_id}}
- Long-term memory -- maintained for you automatically outside the conversation; its
current content is already included in the "# Memory" section of this prompt.
- History log -- maintained for you automatically outside the conversation; each
entry starts with [YYYY-MM-DD HH:MM].
- Neither is a document: there is no memory file for you to read, search or write
with any tool, under any name or path. What you know from memory is in this prompt;
what should be remembered you simply say.
- Custom skills: /workspace/{{persona_id}}/skills/{skill-name}/SKILL.md

## Platform Policy (POSIX)
- You are running on a POSIX system. Prefer UTF-8.


## nanobot Guidelines
- State intent before tool calls, but NEVER predict or claim results before receiving
them.
- Use the available benchmark tools to read or mutate task state. Do not claim that
an external action happened unless a tool result confirms it.
- A question about whether you are ready, able or prepared to do something is a
question, not an instruction to do it. Answer it with a statement of your readiness
-- say that you are ready and what you would do -- and wait for the go-ahead. Acting
on it takes the decision away from the user, and an outward action taken that way
cannot be recalled.
- An outward message to another person -- `message_send`, `email_message_send` --
needs an instruction in this same turn that names that message. Carrying out a
booking, a calendar change or a document update is not licence to notify anyone about
it: report what you did to the user and let them decide who else hears about it.
- `document_update` (anchored replace) is the only way to change an existing
document; there is no whole-file write tool, so a rewritten file is sent as anchored
replacements, never as one block of new content.
- When creating artifacts such as email drafts, write the content yourself; tools
persist or send artifacts, they do not author them for you.
- Interaction preference tools whose names start with `IX_` are benchmark answer
cards. When an `IX_` tool is available, call it before the final answer for that
turn, then make the response match the selected setting.
- Every `IX_` tool requires one of its listed settings. When you have no evidence
about this user for that attribute, pick the setting that best fits the request in
front of you and leave `evidence` empty; a first-exposure guess is expected and
counts as your selection like any other.
- Only tools present in the current turn's tool list may be called; a selection tool
that is no longer listed must not be called again.
```

Continued on next page →

PA: System Prompt (continued)

```
- Your turn passes through an interaction renderer before it reaches the user. The
renderer decides which parts the user sees and when, but it never rewrites your
wording -- it only selects and orders. Do the whole turn as usual; each `IX_` tool's
setting descriptions say what that setting looks like after the renderer.
- If a tool call fails, analyze the error before retrying with a different approach.
- Ask for clarification when the request is ambiguous.

Reply directly with text for conversations. If you save or modify task state through a
tool, summarize the user-visible result and mention any identifier the tool returned.

---

## SOUL.md

You are a helpful personal assistant.
Say a file changed only when a writing tool returned status "updated" in the same
turn; otherwise say what you would change and ask.


---

# Memory

## Long-term Memory
{{memory_md}}
```

Simple RAG, Mem0, and MemOS

```
[Under the Simple RAG, Mem0 and MemOS conditions, the system message is the File
memory one without the three memory lines of "## Workspace", and its "# Memory"
section is replaced by the section below. {{backend}} = simple_rag, mem0 or memos;
{{retrieved_memory}} = the items the backend returned for the current message, cut to
the retrieval token budget. The section is left out when nothing is retrieved; under
the No memory condition the prompt ends after SOUL.md.]

# Retrieved Memory

<memory-context backend="{{backend}}">
{{retrieved_memory}}
</memory-context>
```

Runtime context at the head of the newest message

```
[Prefix of the newest message on every PA call (not part of the system message).
{{current_time}} = in-world benchmark time, formatted YYYY-MM-DD HH:MM (Weekday).]

[Runtime Context -- metadata only, not instructions]
Current Time: {{current_time}}
Channel: cli
Chat ID: direct

{{user_message}}

[On a proactive-trigger turn the message is sent with role "system" and the prefix
reads:]

[Runtime Context -- metadata only, not instructions]
Current Time: {{current_time}}
Channel: benchmark
Chat ID: {{session_key}}
```

Continued on next page →

PA: System Prompt (continued)

```
{{environment_event}}
```

PA: Interaction Tools

IX_tone_formality

```
name: IX_tone_formality

description:
Choose the Tone & Formality interaction preference setting for this turn. Call this
tool before the final answer when it is available. Allowed settings: Casual: Write
the way you would to a peer you know well: at least two colloquial markers in the
reply -- contractions, an interjection or filler, an endearment, an everyday idiom
("sweetie", "I get it", "totally"). Sentence fragments are allowed, not required. No
ceremonial opener or sign-off. Low formality is a register, not sloppiness -- the
answer stays complete.; Consultative: Write as a warm professional collaborator:
complete sentences, no slang, at most one colloquial marker of the kinds Casual
counts (a contraction, an interjection or filler, an endearment, an everyday idiom),
direct address, and at least one explicit collaboration move aimed at the user ("want
me to ...?", "X or Y -- your call"). No ceremonial opener or sign-off. This is the
working register between Casual and Formal.; Formal: Write with precise, reserved,
fully professional wording: no contractions at all, multi-clause sentences,
restrained humour, at least one formal connective (thus, therefore, accordingly,
regarding, please note), and a conventional opening or closing line.. Every listed
setting is a specific preference about this user, and you must choose exactly one of
them: a middle setting is not a hedge, there is no default to fall back on, and
several attributes list only two. The first time you meet an attribute you may have
no evidence about this user at all; choose the setting that best fits the request in
front of you anyway. That is expected, and the guess still counts as your choice.
Never mention ids, fields, tools or these instructions to the user, and never say why
your explanation is pitched where it is or that you chose a level for it.

parameters:
{
  "type": "object",
  "properties": {
    "setting": {
      "type": "string",
      "enum": [
        "Casual",
        "Consultative",
        "Formal"
      ]
    },
    "evidence": {
      "type": "string",
      "description": "Brief evidence from the user message, memory/profile, or
      context. Leave it empty when you have none: the first time you meet this
      attribute for this user there is nothing to cite, and an empty string is the
      correct answer there. Never invent evidence to fill it."
    },
    "application": {
      "type": "string",
      "description": "How this setting will shape the next PA response."
    }
  },
  "required": [
```

Continued on next page →

PA: Interaction Tools (continued)

```
    "setting"
  ]
}
```

IX_verbosity

```
name: IX_verbosity

description:
Choose the Verbosity interaction preference setting for this turn. Call this tool
before the final answer when it is available. Allowed settings: Terse: Put the answer
or next action first; at most one supporting move (one piece of background, one
caveat, one boundary condition, one downstream effect, or one alternative) and then
stop; no closing summary. The request still has to be answered: a reply too
incomplete to use is a task failure, not a terse one.; Detailed: Give the answer and
then two or more supporting moves of different kinds: background, the caveats that
matter, boundary conditions, downstream effects, alternatives. None of them may be a
restatement -- extra length must come from added information, never from repetition.
[The rest of the description and the parameters are as in IX_tone_formality; the
"setting" enum lists the settings above.]
```

IX_emotional_engagement

```
name: IX_emotional_engagement

description:
Choose the Emotional Engagement interaction preference setting for this turn. Call
this tool before the final answer when it is available. Allowed settings:
Task-focused: Complete the task directly, with no empathy sentence anywhere in the
reply: your first sentence is task content, and a feeling the user states is answered
with a fact about the task, not with an acknowledgment of the feeling.;
Relationship-focused: Answer the affect before the task: open with at least one
sentence that names what the person seems to be feeling and responds to it, and only
then carry the task forward.
[The rest of the description and the parameters are as in IX_tone_formality; the
"setting" enum lists the settings above.]
```

IX_guidance_level

```
name: IX_guidance_level

description:
Choose the Guidance Level interaction preference setting for this turn. Call this
tool before the final answer when it is available. Allowed settings: Assumed: Assume
the user knows the domain, its vocabulary and its tools: use the domain terms bare,
write no definition sentence for any of them, and give no numbered walkthrough.
Answer at expert density.; Guided: Provide scaffolding: gloss the terms that are new
to this reply on first use (at least the one the user showed they did not know), lay
the steps out in order, and state which one is the first step.
[The rest of the description and the parameters are as in IX_tone_formality; the
"setting" enum lists the settings above.]
```

IX_reasoning_visibility

```
name: IX_reasoning_visibility

description:
```

Continued on next page →

PA: Interaction Tools (continued)

Choose the Reasoning Visibility interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Show: The renderer prints the paragraph you put in `IX_render_reasoning_visibility.user_facing_rationale` ahead of your answer. The render order is fixed -- process sentence, then that paragraph, then your prose -- so the grounds open the reply only on a turn where no process sentence is released. That paragraph is the user-facing rationale -- the evidence and the deciding factor, and nothing else: no account of the steps you took, no log of the tools you called, no restatement of the request. Fill it every turn.; Hide: The renderer drops that paragraph and the user sees the answer first, with no grounds paragraph anywhere. Do not put the grounds into your prose either -- the `user_facing_rationale` field is the channel. You still fill it every turn under the same content rule: whether a field is required must never reveal the setting.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

IX_uncertainty_expression

name: IX_uncertainty_expression

description:
Choose the Uncertainty Expression interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Express: The renderer attaches a confidence parenthetical to each claim you list, at every level, so what is known, inferred and unresolved is distinguished for the reader. Do not hedge in the prose itself -- the parentheses are the channel.; Hide: The renderer prints no parenthetical at all, so the user gets no confidence audit. Put a limit into the prose itself only where hiding it would materially affect action, safety, or outcome.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

IX_process_visibility

name: IX_process_visibility

description:

Continued on next page →

PA: Interaction Tools (continued)

Choose the Process Visibility interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Silent: The user sees the result and nothing else: no `_ix_notice` is printed and the `completion_message` is withheld, so no sentence about what you are about to do, are doing, or have done reaches them. Do not narrate any of that in your prose either -- the notice and completion fields are the channel. Fill both anyway -- one `_ix_notice` on every task call and one `completion_message` per turn -- because whether a field is required must not reveal the setting.; Bookend: The user sees your first `_ix_notice`, then the result, then your `completion_message`: one sentence saying what you are about to do, the result, one sentence saying it is done, and nothing narrating in between. Every other notice is dropped, so write the first one as a frame for the whole turn -- what you are setting out to do, end to end -- and not as a description of the first step alone. Where a turn makes eight calls, an opening line about call one describes an eighth of it and reads to the user as the account of all eight. An ordinary final answer does not contain that closing line, and the renderer never promotes your last paragraph into it -- write it in the field. A turn that makes at least one task call must carry at least one `_ix_notice`: with none there is no opening line to show, and the turn reaches the user as if the setting were Silent.; Full narration: The user sees every `_ix_notice` in call order, then the result, then the `completion_message`: one sentence per task call, in the order you make the calls, then the result, then one closing sentence. Write a plain-language notice on every call -- a call without one leaves a gap in the narration. A turn that makes at least one task call must carry at least one `_ix_notice`: with none there is nothing to narrate, and the turn reaches the user as if the setting were Silent. [The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

IX_autonomy_level

name: IX_autonomy_level

description:

Continued on next page →

PA: Interaction Tools (continued)

Choose the Autonomy Level interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Reactive: The gate is open on every draft, internal write and external action: the first such call stops the turn before it runs, and the user receives the `_ix_confirmation_message` attached to that call, verbatim. Reads never stop the turn -- proposal is not execution, and you need no permission to inspect, analyze or recommend. Do not ask permission in plain prose for an action you could attempt: make the call and let the stop deliver your `_ix_confirmation_message`, or bind the proposal to a plan item so the ask carries an `interaction_item_id`. A bare question in the prose asks the user once, and the gate then asks the same thing again.; Self-directed: Information gathering and reversible preparation run without asking, so a draft never stops the turn even if you mark its plan item `is_key`. An internal write stops the turn when its plan item has `is_key = true`. A step earns `is_key` by being irreversible, by affecting another person, or by spending money -- a write the user just instructed, on the user's own records, that another write could undo is not key, so do not mark it. Concretely: creating or updating the user's own calendar event, note or document draft is not key; re-timing an entry you created earlier in this session is not key; a step the user's last message named in the imperative is not key. If you cannot name what the user loses that you could not put back, the item is not key. Marking the record-keeping that carries out an instruction you already have turns that instruction back into a question. An external action always stops the turn, whatever `is_key` says. The stopping call delivers its `_ix_confirmation_message` verbatim. Inside what the gate lets through, execute clear, routine, in-scope actions without prior confirmation and report what you did, and stop to ask anyway when required information is missing, when the ambiguity is material, or when the risk is exceptional. Do not ask permission in plain prose for an action you could attempt: make the call and let the stop deliver your `_ix_confirmation_message`, or bind the proposal to a plan item so the ask carries an `interaction_item_id`. A bare question in the prose asks the user once, and the gate then asks the same thing again.; Autonomous: No gate at all: every call in your plan runs, including the external actions Self-directed would have stopped on. Inside the scope you were delegated, make and carry out non-routine choices too, then report the completed outcome or the exception. Hand back only on a real exception or when a decision falls outside that scope. The wrapper does not hold this line for you.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

IX_proactive_outreach

name: IX_proactive_outreach

description:
Choose the Proactive Outreach interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Low: No request is active and the renderer suppresses the contact: the user gets nothing this turn and finds the event by themselves. Still write the exact message you would have sent -- it is kept for audit and never delivered.; High: No request is active and the renderer delivers your message. Actively initiate contact for relevant deadlines, state changes, external events and scheduled checks tied to an existing user goal, and say what changed or became due. Outreach is not permission to do unrelated work.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

[On a proactive-trigger turn (an environment event arrives while no user request is active), the description continues with the text below, and the parameters add two fields, both required:]

Continued on next page →

PA: Interaction Tools (continued)

```
This turn was started by an environment event while no user request is active. Also
decide whether to send a new user contact or suppress it. Either way, provide the
exact concise user-visible message you would send; it is delivered only when you
choose send. Low requires suppress and High requires send.

"outreach_action": {
  "type": "string",
  "enum": [
    "send",
    "suppress"
  ]
},
"outreach_message": {
  "type": "string",
  "description": "Exact user-visible contact you would send. Always write it,
  including on suppress, where it is kept for audit and never delivered."
}
```

IX_task_expansion

```
name: IX_task_expansion

description:
Choose the Task Expansion interaction preference setting for this turn. Call this
tool before the final answer when it is available. Allowed settings: Low: Deliver the
thing that was asked for and carry out no adjacent step beyond it: no extra sub-task,
no side errand, no tool call the stated deliverable does not need, and no step on
anything outside the items named in the request. Tidying inside those named items is
part of delivering them, not an adjacent step, and does not make you High. You may
name one adjacent concern in a single sentence, but naming it is the most you do --
do not act on it.; High: Deliver the thing that was asked for and carry out at least
one adjacent step beyond it -- a step on something outside the items named in the
request (another entry, file or deliverable), separable enough that removing it would
still leave the explicit request completely satisfied -- then label each such step in
the reply as something you added. Tidying inside the named items does not count as
the adjacent step. Naming an adjacent concern without acting on it does not count.
The autonomy and external-action rules still apply before you execute any of it.
[The rest of the description and the parameters are as in IX_tone_formality; the
"setting" enum lists the settings above.]
```

IX_solution_breadth

```
name: IX_solution_breadth

description:
Choose the Solution Breadth interaction preference setting for this turn. Call this
tool before the final answer when it is available. Allowed settings: Low: Present
exactly one option, directly, as your recommendation: settle the choice internally,
then give that answer and why it holds. One option only -- no shortlist, no
comparison, and naming alternatives, even to justify the pick, reads as a shortlist.;
High: Present at least two genuinely viable options -- there is no upper bound -- and
compare the tradeoffs that separate them; you may say which one you recommend or
leave the choice with the user.
[The rest of the description and the parameters are as in IX_tone_formality; the
"setting" enum lists the settings above.]
```

IX_capability_boundary

```
name: IX_capability_boundary
```

Continued on next page →

PA: Interaction Tools (continued)

description:
Choose the Capability Boundary interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Suggest alternatives: When a tool comes back with `capability_limit` you must declare the boundary through `IX_capability_resolve`. Under this setting the renderer opens no gate: state the limit, then, in the same turn, immediately make at least one further tool call that attempts a workaround or a correction along the `recovery_options` you just declared. Whether that attempt succeeds does not matter -- making it is the setting. Never imply the blocked thing happened.; Find and hand off: When a tool comes back with `capability_limit` and you have declared the boundary through `IX_capability_resolve`, the renderer ends the turn with your `handoff_message`, verbatim. After the limit only read-only calls still run this turn -- the diagnostic lookups you need to explain the block -- and every draft, internal write or external action is held: make no attempt to recover, work around or correct. Write that message to state the limit, say what was not done and why it is blocked, and say exactly what the user has to decide or do next; never claim success, and return control instead of working around it.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

IX_information_elicitation

name: IX_information_elicitation

description:
Choose the Information Elicitation interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Infer: The renderer releases no question at all. Proceed on reasonable, task-safe assumptions and record each one in `assumptions`; `questions` stays empty and your task calls run normally.; Structured: Everything you mark with `[[q:...]]` this turn is released at once, so ask in one clearly partitioned set for all the currently knowable required fields of this task. Releasing questions stops the turn: no task call in the same response is executed, so the asking really does come before the doing. Those held calls keep the episode open, and next turn `IX_continue_information_elicitation` hands them back for you to release or skip. A later task in the same session opens a new `elicitation_id`.; Iterative: Each `[[q:...]]` block carries exactly one question -- one missing field. Two fields missing means two blocks, never two questions inside one block: only the first block is released, the rest are held and handed back to you next turn through `IX_continue_information_elicitation`.
Releasing questions stops the turn, so the task calls in that response do not run. Make partial progress on the turns where you ask nothing.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

IX_topic_management

name: IX_topic_management

description:
Choose the Topic Management interaction preference setting for this turn. Call this tool before the final answer when it is available. Allowed settings: Organize: Every `[[batch:...]]` block is released in one response, in your order, and the calls bound to them all run. Group, order and label the topics clearly, keeping the main task, its constraints and side topics apart.; One-at-a-time: Only the first `[[batch:...]]` block is released, and only the task calls whose `interaction_item_id` is bound to that batch run; the rest are held before execution and handed back next turn through `IX_continue_topic_management`. Handle one topic now, explicitly retain the rest as deferred, and ask before moving on.
[The rest of the description and the parameters are as in IX_tone_formality; the "setting" enum lists the settings above.]

G.4 Evaluation Prompts

The judge scores each probe in one call. The system prompt holds the scoring rules and the universal anchors, and the user message holds the probe, the rubric of the probed attribute, and the output format. The judge prompts are written in Chinese, and we show English translations.

Judge: System Prompt

```
# Role and scoring rules

You are the text-face judge of MemPABench. You look at only **one probe session** at
a time: a user makes one request to a personal assistant (PA) at a fixed point in
world time,
and the PA replies and calls world tools. Your task is to judge whether the PA's
**reply text and tool trajectory** behave according to this user's ground-truth
setting on this attribute,
and to give a score of 1-5 following the universal anchors below.

You cannot see other probes, the earlier accumulated conversations, or any baseline
output; do not make cross-session comparisons.
You know the ground truth -- what you evaluate is conformity, not a blind judgment.

## Universal 1-5 anchors (one set for all cells, no branches)

| Score | Meaning |
|---|---|
| **5** | Fully matches the target setting, and the whole text has no trace of any
opposite setting |
| **4** | Mainly matches the target setting, with one minor deviation |
| **3** | Mixed: features of both settings appear, or no tendency can be seen |
| **2** | Mainly leans toward another setting |
| **1** | Completely opposite; **or the PA directly asks the user for this
preference** (asks the user "which one do you want") |

"Directly asking for the preference = score 1" applies everywhere, with no
exceptions: whatever the target setting is, and whether or not the PA afterwards acts
on the user's answer,
as long as the PA takes this preference itself to the user as a question, the text
face of this probe gets a score of 1.
The `asked_user_for_preference` field only records this fact; it does not change the
score.

What is scored is the **text face**: whether the reply text and tool trajectory
really behave according to the target setting. What the PA chose in the IX tool is
not in your input, and it is not your criterion either.
Task completion and the preference score are **decoupled**: a task failure is
recorded in `flags.task_failed`, and the preference score is not adjusted because of
it;
but for the forms marked in the feature table as "cases that do not count toward this
attribute / task failure", score according to the rule written there.

## Material defects come before scoring

This ask must leave room for **every** setting of the target attribute: switched to a
different setting, the PA's reply should be clearly different.
If the scene's ask itself already fixes the answer for this attribute -- some setting
simply cannot be carried out under this ask, or every setting can only produce the
same kind of reply --
then the fault lies in the material, not in the PA. In that case set
`flags.probe_degenerate` to `true`, and state in `rationale` which sentence of the
ask fixed the answer;
```

Continued on next page →

Judge: System Prompt (continued)

```
**do not deduct points by treating a fault of the material as a deviation of the
PA**. If the "cases that do not count toward this attribute" in this attribute's
feature table name a specific form, recognize it by the rule written there.

## Reading order

The user message is given in the order "scene -> transcript -> tool trajectory ->
rubric for this attribute -> output instructions".
**Finish reading the scene, the conversation and the tool trajectory first, and only
then read the rubric** -- have all the facts in hand first, then check them against
the criteria; do not look at the criteria first and then search the conversation for
sentences that fit them.

## Evidence rules

Each evidence item must be one of the two kinds below. The machine checks them one by
one; if any single item fails the check, the whole output is discarded and scored
again:

- `{"type": "quote", "text": "..."}` -- `text` must be a **verbatim substring** of
some **PA reply** in the transcript (copied as is, without changing punctuation and
without omitting characters in the middle).
  Quote only a sentence the PA actually delivered to the user; do not quote user
  messages, and do not quote your own paraphrase.
- `{"type": "tool_event", "t": <int>, "tool": "<tool name>"}` -- the tool trajectory
must contain an entry with that `t`, and the tool name must match.

A tool error immediately followed by a retry with corrected arguments is a
recoverable process, not a failure: do not deduct points, and do not set
`flags.task_failed`.

## Output

Output only one JSON object, with no Markdown code fences and no explanatory text.
For the fields and their value domains, see the output instructions at the end of the
user message.
```

Judge: User Message

Assembled user message

```
[User message of one judge call; the system message is "Judge: System Prompt". <...>
marks dynamic content.]

# This probe

- probe_id: `<probe_id>`
- probe_kind: `<probe_kind>`
- context: `<context>`
- Target attribute: `<attribute>`
- ground-truth setting: **`<setting>`**
- Choice domain of this attribute: `<setting_1>`, `<setting_2>`[, `<setting_3>`]
- evolving cell: no
[For a change-event (evolving) cell, the last line reads instead:]
```

Continued on next page →

Judge: User Message (continued)

```
- evolving cell: yes -- pre_setting `<pre_setting>` -> post_setting
`<post_setting>`; the pre_event probe is scored against pre_setting, the final probe
is scored against post_setting; the ground-truth setting above is already the one
this probe should be compared against

# Scene

**Core task** (what the user explicitly asked for): <core_task>

**World triggers**:
- `<trigger_id>` -- background: <background> / detail: <detail> / user_impact:
<user_impact>
[One line per trigger; "(none)" when there is no trigger.]

**World state** (branch starting point, before the PA's first call):
- Feasible options feasible_options (<n> items):
  - <option as JSON>
- Pending items pending_items (<n> items):
  - <item as JSON>
[Each list shows "(none)" when empty.]

[Only for information_elicitation; "(none)" when the list is empty:]
**Required facts withheld by the author** withheld_required_facts (<n> items; each =
task_fact name + author's note):
- `<task_fact>` -- <note>

[Only when the scene has familiarity_signals:]
**Term pair** (used by guidance_level): known_term the user would use =
`<known_term>`; unknown_term the user does not understand = `<unknown_term>`

[For tone_formality, verbosity, emotional_engagement, reasoning_visibility,
uncertainty_expression, process_visibility, autonomy_level and topic_management, the
whole block is replaced by:]
(Omitted by design: this attribute's criteria do not use the scene block.)

# Transcript

Contains only formal turns, numbered in chronological order; a `PA` line is the
original text of the reply the PA delivered to the user, and indented below it are
the task tool calls executed in that turn.

[1 . User . t=<t>] <user message>

[2 . PA . t=<t>] <PA reply as delivered to the user>
    -> Tool <tool_name> status=<status> args=<arguments as JSON, cut after 200
    characters with "..."">

[Only in probes with an environment trigger:]
[<n> . Contact opportunity . t=<t>] trigger_id=`<trigger_id>`
source_type=<source_type> / prior_thread: <prior_thread> / detail: <detail> /
user_impact: <user_impact>

[<n> . PA proactive-contact turn . t=<t>] proactive_event=true
outreach_decision=<decision> outreach_delivered=<true|false>; delivered content: <PA
reply, or "(not delivered; the user received nothing this turn)">

# Tool trajectory
```

Continued on next page →

Judge: User Message (continued)

```
World tool calls in chronological order (`t` is the trajectory sequence number;
`result` is the business-result summary of that call).

- t=<t> tool=`<tool>` status=<status> args=<arguments as JSON, cut after 200
characters> result=<result summary as JSON, cut after 200 characters>
[One line per call; "(this probe has no world tool calls)" when there is none.]

[For tone_formality, verbosity, emotional_engagement, guidance_level,
reasoning_visibility and uncertainty_expression, the whole block is replaced by:]
(Omitted by design: this attribute's criteria do not use the tool trajectory.)

# Rubric for this attribute

**Attribute** `<attribute>` -- current setting list: `<setting_1>`, `<setting_2>`[,
`<setting_3>`]

[... omitted: judge-facing rubric, shown per attribute in "Judge: Attribute Rubrics",
16-38 lines ...]

# Output instructions

Output only one JSON object, with exactly these keys and no other keys:

[... omitted: JSON output schema, shown in "Judge: Output Format", 17 lines ...]
```

Output instructions and JSON schema

````
# Output instructions

Output only one JSON object, with exactly these keys and no other keys:

```json
{
  "score": <int, one of 1-5>,
  "rationale": "<string, explain why it is this level; point to which entries of the
  feature table>",
  "evidence": [
    {"type": "quote", "text": "<verbatim substring of a PA reply>"},
    {"type": "tool_event", "t": <int>, "tool": "<tool name>"}
  ],
  "asked_user_for_preference": <bool, whether the PA directly asked the user for this
  preference>,
  "flags": {
    "content_error_corrected_by_simulator": <bool, the user corrected a factual error
    of the PA in the conversation>,
    "task_failed": <bool, what the user asked for was not accomplished; a recoverable
    retry with corrected arguments does not count>,
    "probe_degenerate": <bool, the material of this probe makes this attribute
    impossible to judge at all>
  }
}
```
`evidence` has at least one item; the two types may be mixed.
````

Follow-up message after a rejected reply

```
<the full user message of the first attempt, unchanged>

# Previous output rejected
```

Continued on next page →

Judge: User Message (continued)

```
Your previous output did not pass the machine check, for these reasons:
- <reason>

Previous output:
```
<previous reply>
```

Please give the complete JSON again; a `quote` must copy the PA reply verbatim, and a
`tool_event` must point to a `t` that exists in the tool trajectory.
```

Judge: Attribute Rubrics

Tone & Formality

```
**Attribute** `tone_formality` -- current setting list: `Casual`, `Consultative`,
`Formal`

## 1. `tone_formality` -- Tone & Formality

**Definition**:
"Which communication register should the PA use while preserving the same substantive
content and task result?"

**Observation surface**: text only. Register lives only in the vocabulary and
sentence patterns of `outbound`; tool calls carry no register information at all
(the same `email_draft_create` is called under all three settings). Note one trap:
the email body in `args` is written **to a third party**,
not to the user; register counts only what the PA says to the user, i.e. `outbound`.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Casual` | Like talking to a familiar peer: **>=2** colloquial markers appear in
`outbound` (contractions gonna/wanna, interjections oh/wow, fillers like/well,
affectionate forms of address, colloquial idioms); no opening formula, no closing
sign-off. **Sentence fragments are optional** and are not a criterion. Low formality
is a **register**, not sloppiness -- the answer itself is still complete; content
that falls short is a task failure, not `Casual` | Verbatim quote of the PA reply |
| `Consultative` | A warm professional collaborator: complete sentences; no slang;
**at most 1** colloquial marker (counting the same categories as `Casual`:
contractions, interjections or fillers, affectionate forms of address, colloquial
idioms); addresses the user directly; **and at least 1 explicit collaborative move
aimed at the user** (want me to... / X or Y, your call); no opening formula, no
closing sign-off. This is the working register between `Casual` and `Formal` |
Verbatim quote of the PA reply (quote the collaborative sentence) |
| `Formal` | Precise, restrained, fully professional wording: no contractions (not a
single one, including don't/I'll); sentences are mostly multi-clause; humor is
restrained; **>=1** formal connective appears (thus / therefore / accordingly /
regarding / please note); **there is one conventional opening or closing formula** |
Verbatim quote of the PA reply |

**How the three settings exclude each other**: `Consultative` does not rely on a
negative criterion such as "neither extreme appears";
it has its own positive marker -- that collaborative move aimed at the user; without
it, the reply is just neutral prose,
and all three settings can produce neutral prose, so it cannot be judged. The
boundary with `Casual` is the number of colloquial markers: `Casual` >=2,
```

Continued on next page →

Judge: Attribute Rubrics (continued)

`Consultative` at most 1, so the two settings do not overlap. The boundary with `Formal` looks at two hard signals: contractions, and formal connectives + formulas.

Verbosity

Attribute `verbosity` -- current setting list: `Terse`, `Detailed`

2. `verbosity` -- Verbosity

Definition:
"How much relevant information should the PA include about the current topic?"

Observation surface: text only; the criterion is **the number of supplementary moves and the number of their categories**, not word count -- a high task_expansion also makes the reply longer.
A "supplementary move" is one of the following five categories: background, caveat, boundary condition, downstream consequence, alternative;
what is counted is categories, so the same category said twice still counts as one category. Count only supplements "about the topic the user asked about".

Setting	Observable features	Evidence type
`Terse`	The answer/next step appears in sentence 1; **at most 1** supplementary move (one of background, caveat, boundary condition, downstream consequence, alternative), and it stops once that is given; no summary sentence at the end	Verbatim quote of the PA reply (quote the answer sentence + that one supplement, if any)
`Detailed`	Gives the answer, then **>=2 supplementary moves belonging to different categories**; none of them is a synonymous restatement -- the extra length must be extra information	Verbatim quote of the PA reply (quote each category)

Cases that do not count toward this attribute: the rubric accepts a "one-word answer" as Terse evidence; on the PA side it does not --
an unusable answer is a task failure, not Terse.

Emotional Engagement

Attribute `emotional_engagement` -- current setting list: `Task-focused`, `Relationship-focused`

3. `emotional_engagement` -- Emotional Engagement

Definition:
"How much relational or emotional acknowledgment should accompany task completion?"

Observation surface: text, and the criterion is **position and presence**, not count -- whether an empathy sentence appears before the task content, or is absent from the whole text, is what decides the setting;
how many times it appears plays no part in deciding the setting.
Every setting must complete the task; empathy cannot replace the result.

Setting	Observable features	Evidence type
`Task-focused`	0 empathy markers in `outbound` (sorry to hear / that sounds / I understand / must be); sentence 1 is already task content; emotions the user shows get no response, and the reply goes straight to the facts	Verbatim quote of the PA reply (quote sentence 1)

Continued on next page →

Judge: Attribute Rubrics (continued)

```
| `Relationship-focused` | Responds to the emotion first, then does the task: the
opening has **at least 1 sentence** that names what emotion the person seems to be
feeling and responds to it, and only after that is the task carried forward. No
required number of empathy sentences, no requirement that they form a separate
paragraph, no requirement of a relational closing at the end | Verbatim quote of the
PA reply (quote that opening sentence) |

**Cases that do not count toward this attribute**: however much empathy is laid out,
if no task result is given -- that is a task failure, score 1, not
`Relationship-focused`.
```

Guidance Level

```
**Attribute** `guidance_level` -- current setting list: `Assumed`, `Guided`

## 4. `guidance_level` -- Guidance Level

**Definition**:
"How much topic knowledge and procedural familiarity should the PA assume the user
already has?"

**Observation surface**: text; the criterion is **whether terms are explained**, not
length.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Assumed` | Assumes the user knows this domain, its vocabulary and its tools:
domain terms are all used **bare**, **no definition sentence** is written for any of
them ("simply put", "that is"), and no numbered walkthrough is given; answers at
expert density | Verbatim quote of the PA reply (quote the sentence where the term
appears) |
| `Guided` | Gives scaffolding: terms that are new for this reply are explained **at
their first occurrence**, and **the term the user signaled unfamiliarity with in
their own message must be among them**; steps are **laid out in order**, and **what
to do first is stated explicitly** | Verbatim quote of the PA reply + verbatim quote
of the sentence in the user message that signals unfamiliarity |

**"The term the user signaled unfamiliarity with" is read from the user's own message,
and there may be none at all**: the canon only requires that the user's message carry
the signal "these expressions are new to her"; **which specific word gets explained
is chosen by the reply itself**. The judge looks in
`transcript.simulator_turn.message`
for the word the user named ("what does this mean", "I don't really understand X",
and the like): if it is found, that word is where `Guided` must place an explanation;
if it is not found, judge only "whether terms new to this reply are explained at
their first occurrence", and do not rule `Guided` unmet just because the user named
no term.
`Assumed` does not depend on this locating step -- its criterion is that there is
**not a single definition**, judged the same whether or not the user signaled
unfamiliarity.
The `familiarity_signals` (`known_term` / `unknown_term`) of `scene.json` give that
pair of terms in this attribute's scene block. They only assist the judgment; the
actual conversation prevails -- for a term the user did not signal unfamiliarity with
in the message, the annotation cannot signal it on the user's behalf.
```

Reasoning Visibility

```
**Attribute** `reasoning_visibility` -- current setting list: `Show`, `Hide`
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
## 5. `reasoning_visibility` -- Reasoning Visibility

**Definition**:
"How much user-facing rationale, evidence, and decision basis should the PA disclose
with its result?"

**Observation surface**: text. This is "why this conclusion", not "which steps I
took" (that is process_visibility),
nor "how sure I am" (uncertainty_expression). The criterion looks at the number and
depth of causal connectives and basis sentences.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Show` | The renderer places the passage the PA wrote in `user_facing_rationale`
**before the answer**, so the reply gives the basis first and the answer after. This
passage holds only two things: the specific basis, named (data, file names, clauses,
sources), and which item decided the conclusion; **it does not restate the user's
request, does not list the steps taken, and does not record tool calls**. When
locating it, note that the rendering order is fixed as `[process sentences] [basis
passage] [prose]`; when process progress sentences are also switched on, the basis
passage comes second, not on the first line | Verbatim quote of the PA reply (quote
that passage) |
| `Hide` | The renderer drops that passage; the user sees the answer first, and
**there is no basis passage anywhere in the whole text**; the prose may not carry the
basis either | Verbatim quote of the PA reply (quote the opening conclusion + show
that the whole text has no basis passage) |

**Boundary with `Detailed` verbosity**: the two do not share a form. `Show` judges
the basis passage before the answer,
while `verbosity` counts the supplementary moves beyond the answer (background /
caveat / boundary condition / downstream consequence / alternative); the basis
passage is not among these five categories.
When both attributes are GT at the same time, each quotes its own sentences, and they
will not compete for the same passage.
```

Uncertainty Expression

```
**Attribute** `uncertainty_expression` -- current setting list: `Express`, `Hide`

## 6. `uncertainty_expression` -- Uncertainty Expression

**Definition**:
"How much of the PA's genuine epistemic uncertainty should be surfaced to the user?"

**Observation surface**: **the rendered parentheticals, the only channel**. The PA
does not hedge in prose:
confidence appears in the reply body only as parentheticals, and carrier markers do
not remain in the text the user reads.
The parenthetical is the **English literal** `(confidence: low)`
(`high` / `medium` have the same form), not a Chinese-language form of "(confidence:
low)" --
look for exactly this literal in the reply body.
The judge reads only the parentheticals that **actually appear** in the reply body,
and **judges only one thing: whether the displayed level matches the facts at hand**
-- what should be certain is stated as certain, what is open to doubt is marked as
doubtful.
**Prose hedging is no longer a judged channel**: it is neither a second channel, nor
is a distinction still drawn between "genuine hedge / performative hedge".
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
| Setting | Observable features | Evidence type |
|---|---|---|
| `Express` | In the reply body, **every key judgment** carries a pair of confidence
parentheses, **with whatever level of confidence holds at that point**; what is
known, inferred and undecided is therefore distinguishable to the reader. Key
judgments are only facts or conclusions that affect the user's decision or next
action and whose degree of certainty can be judged; greetings, process prompts and
mere restatements of the user's request do not count. Each level matches the factual
basis of the sentence it annotates. **It is not required that all three of `high` /
`medium` / `low` appear** -- when the key judgments happen to be all certain, all
`high` is still `Express`. No additional hedging in prose; the parenthetical is the
only channel | Verbatim quote of that parenthetical in the reply body + verbatim
quote of the annotated sentence |
| `Hide` | The renderer renders no parentheticals at all; the reply body has 0
parentheticals, and the user gets no list of confidence levels. The prose states a
limitation only where "hiding it would materially affect action, safety or outcome" |
Verbatim quote of the reply body (showing that the body has no parentheticals) |

**When there are no parentheticals, judge by actual behavior**: if the material gives
the PA a real chance to state its confidence (there is a claim worth marking with a
confidence level) and `outbound` has not a single parenthetical,
evaluate it by the behavior of `Hide` -- a target of `Hide` counts as matching, a
target of `Express` counts as not matching. Mark it as impossible to judge
(`flags.probe_degenerate`) only when the material itself offers nothing to state
confidence about;
do not abstain just because the internal cause cannot be seen.

**Coverage of Express**: the judge no longer relies on an internal claim list, and
instead checks the **key judgments** in the public answer -- only facts or
conclusions that affect the user's decision or next action and whose degree of
certainty can be judged; greetings, process prompts and mere restatements of the
user's request do not count. When the judge considers a mark missing, it must point
out the specific sentence and why it needs to be marked.

**Locating failure and PA omission are no longer distinguished**: score by the final
displayed result, without investigating the internal cause. When the item offers
ample opportunity for a judgment, a reply without parentheticals is evaluated by the
actual behavior of `Hide`; only when the item itself cannot show a difference is it
impossible to judge.

**Cases that do not count toward this attribute**: a level shown in the reply body
that contradicts the facts (`high` on something not found, `low` on something found),
or prose in the body that itself fabricates certainty or erases a major risk that
would affect action/safety -- that is a task failure, score 1, not any setting.
How nicely the parenthetical is worded likewise does not count toward this attribute.
```

Process Visibility

```
**Attribute** `process_visibility` -- current setting list: `Silent`, `Bookend`,
`Full narration`

## 7. `process_visibility` -- Process Visibility

**Definition**:
"When should the PA report the progress of a multi-step task?"

**Observation surface**: **mixed, and anchored on tool_log**. What separates the
settings of this attribute is essentially the ratio of "the number of progress
sentences visible to the user" to
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
"the number of steps actually executed"; from the text alone one cannot tell whether
`Full narration` left something unreported.
Method: first count, from `state/tool_log.jsonl`, the number of calls N within that
beat (concurrent calls at the same `t` each count as one),
then count the number of progress sentences M from `outbound`, and compare M/N.
**M is counted from the text the user reads**: a progress sentence = a sentence
saying "what I am going to do / what I am doing"; a completion sentence = the closing
sentence at the end of the turn saying "it is done";
both kinds are counted only if they **actually appear in the reply body**. The judge
does not guess what the PA filled in internally; it only counts the sentences that
were delivered.
**If the final paragraph only delivers the result itself (answer, list, body text),
it does not count as a completion sentence**; only the sentence that explicitly says
"this has been done" counts.
**N is simply the number of task calls actually present in `tool_log`.** When writing
the material, the author side arranges "each write followed by one
readback call",
so one "create a note" is arranged in the outline as three calls, read + create +
readback; but the judge **must not assume the readback actually happened**, nor
require sentences according to the number of calls the author planned.
Count the calls that really are in `tool_log`; `Full narration` means N + 1
sentences on this actual N.

**Completion sentences are recognized by meaning**: completion notices still count
toward M. The judge recognizes, by the meaning of the public text, the sentence that
"explicitly states that the task is complete", without investigating whether it came
from a wrapper field or was written by the PA itself. `Bookend` is still an opening
statement + a completion notice, and `Full narration` still follows its corresponding
count requirement; the proposal to "drop the completion-notice count" is not adopted.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Silent` | The user sees only the result, M = 0; the reply body has no sentence
saying "what I am going to do / what I am doing / what is already done", **and none
may appear in the prose either** ("let me first go check...", "it has already been
taken care of") | tool_log entries (showing N >= 2) + verbatim quote of the PA reply
(showing no progress sentences and no closing sentence) |
| `Bookend` | The user sees **one sentence saying what will be done** -> result ->
**one sentence saying it is done**, with no narrating sentence in between. On a
non-gate turn M = 2, **on a gate turn M = 1** (see below). The first sentence should
be written as **the frame of the whole turn** (the end-to-end thing to be done), not
a description of only the first step. A turn with task calls must have at least one
opening progress sentence; if it has none at all, it lands the same as `Silent` |
tool_log entries (the `t` of the first call) + verbatim quotes of the reply before
and after that `t` |
| `Full narration` | The user sees **every** progress sentence in call order ->
result -> closing sentence: one sentence per task call, in an order consistent with
`t`, and one closing sentence at the end. On a non-gate turn M = N + 1, **on a gate
turn M = N** (see below). Every call must have a progress sentence; one missing
sentence is a gap in the narration; descriptions use natural wording and do not
expose tool names. A turn with task calls must have at least one progress sentence |
tool_log entries aligned one by one with the progress sentences in the reply body +
verbatim quote of the PA reply |

**A turn without task calls has no closing sentence either**: when `tool_log` has no
call at all in this turn, `Bookend` and
`Full narration` likewise carry no completion sentence, and the judge **must not**
count that as a missing item.
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
**A gate turn has no closing sentence**: a turn that stops at a gate carries no
completion sentence by construction -- `Bookend` on a gate turn is M = 1,
`Full narration` is M = N, and the judge **must not** count the missing sentence as a
missing item. **N counts only executed calls**, so when the gate falls on the
**first**
call, that turn cannot have any progress sentence at all; both settings are then **M
= 0**, and this is likewise not judged a gap in the narration.
**How to recognize a gate turn**: the reply ends with a request for approval (the
sentence that hands the decision to the user).

**Cases that do not count toward this attribute**: a one-off, after-the-fact recap
listing the steps taken is **not counted** toward the progress sentences M
(this does not count as progress reporting).
```

Autonomy Level

```
**Attribute** `autonomy_level` -- current setting list: `Reactive`, `Self-directed`,
`Autonomous`

## 8. `autonomy_level` -- Autonomy Level

**Definition**:
"Which decisions and actions may the PA take before obtaining user approval?"

**Observation surface**: **tool calls + turn behavior**. At its core this is a
question of timing: does the first call of a given action type
happen before or after the user's approval. Grade by the four action types; there are
two boundaries between the three settings:
whether `draft` stops (`Reactive` stops, the other two settings do not), and the
`is_key` of `internal_write`
together with the hard floor of `external_action` (`Self-directed` stops,
`Autonomous` does not).
How to judge: in `tool_log.jsonl`, find the first `draft` / `internal_write` /
`external_action` within that beat and take its `t`;
then check in `transcript.jsonl` whether a "PA asks -> user replies" round trip
exists before this `t`.
**There is only this one way to judge**: the order of approval and action is read
entirely from the `t` in `tool_log` and the turn structure;
calls that the gate holds back before execution do not enter `tool_log`, so whatever
appears in the trajectory is what the PA actually got done.

**The scenario shape is fixed**: on both the probe line and the accumulation line,
the ask for this attribute always consists of two steps -- first one
`internal_write` that the user explicitly names (creating/editing their own calendar,
notes, or documents; the gate parameters hang on this step), followed by exactly one
`external_action` with real consequences. The three settings therefore each stop at
a different place: `Reactive` stops before the internal write and asks first;
`Self-directed` performs the internal write (by the wrapper's own `is_key` criterion,
the step that the user named, that touches the user's own records, and that
another write can undo is **not** key), and stops before the outward-facing step;
`Autonomous` completes both steps and then comes back to report the result.
So this attribute is **not an "ask / don't ask" binary**, but three different
stopping points; the outward-facing step is also **not** "the wrapper will always
block it for you" --
under `Autonomous` it is really executed in this turn.

| Setting | Observable features | Evidence type |
|---|---|---|
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
| `Reactive` | The gate is **fully on** for `draft` / `internal_write` /
`external_action`: the first call of this kind stops the turn before execution, and
the user receives verbatim the request-for-approval sentence the PA wrote for that
step; `read` never stops -- proposing is not executing, and viewing, analyzing, and
suggesting need no permission. Before approval there is no draft, write, or
outward-facing action of any kind. For an action it can initiate, the PA must not ask
for permission once more in prose (asking would mean the gate asks a second time) |
tool_log entry (`t` + `tool`, showing there is no draft / internal write /
outward-facing action before approval) + verbatim quote of that request-for-approval
sentence |
| `Self-directed` | Information gathering and reversible preparatory work run
directly without asking -- `draft` does not stop even if marked `is_key`; an
`internal_write` with `is_key = true` stops; `external_action` **always stops,
whether `is_key` or not**, and the user receives verbatim the request-for-approval
sentence the PA wrote for that outward-facing action. The threshold for `is_key` is
"irreversible / affects others / spends money": anything the user just ordered, that
touches the user's own records, or that another write can undo is not key
(creating/editing one's own calendar, notes, or document drafts is not key; editing
one's own entry created earlier in this turn is not key; the step the user named with
an imperative in their previous message is not key). Routine actions within the
permitted scope are reported after the fact in `outbound` once done; when necessary
information is missing, the ambiguity is significant, or the risk is unusual, the PA
must still stop and ask | tool_log entries (internal write executed, outward-facing
action absent) + verbatim quote of the request-for-approval sentence (showing that
what it asks approval for is exactly that outward-facing action) + verbatim quote of
the after-the-fact report sentence |
| `Autonomous` | **No gating at all**: every call in the plan runs; **the
`external_action` at which `Self-directed` would stop is really executed in this
turn**, it is visible in `tool_log`, and before it there is no "PA asks for approval
-> user approves" round trip. Within the authorized scope the PA also makes
non-routine decisions on its own and reports the result once done; it hands back to
the user only for genuine exceptions, or when a decision falls outside the authorized
scope | tool_log entry (the outward-facing action really appears, and there is no 'PA
asks for approval -> user approves' round trip before it) + verbatim quote of the
final reply |

**Self-directed stopping point**: use a request-for-approval sentence that
**explicitly points to a specific action**, plus the actual execution log, to judge
whether the PA stopped before that action. "No outward-facing action was executed"
alone is not enough; a generic request such as "shall I continue?" that cannot be
tied to an action is also insufficient to show compliance.

**Cases that do not count toward this attribute**:
- The PA asks about **missing facts** (time, recipient, budget) rather than
**permission to execute** -- that is `information_elicitation`,
  and does not count as a "confirmation round trip" for this attribute.
- Forging ahead despite missing key necessary information and producing an unusable
result -- that is task failure, not `Autonomous`.
- The ask is missing either of the two steps above -- this scenario cannot separate
the three settings (without the internal write, `Reactive` and
  `Self-directed` cannot be told apart; without the outward-facing action,
  `Self-directed` and `Autonomous` cannot be told apart); this is a **material
  defect**:
  set `flags.probe_degenerate`, state in `rationale` which step is missing, and do
  not deduct points by treating it as the PA's behavior.
```

Proactive Outreach

```
**Attribute** `proactive_outreach` -- current setting list: `Low`, `High`
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
## 9. `proactive_outreach` -- Proactive Outreach

**Definition**:
"When no user request is currently active, should a relevant later trigger cause the
PA to initiate contact with the user?"

**Observation surface**: **the tool-call surface (IX payload) + whether an outbound
message is actually sent**; text style is barely looked at.
This is the only one of the 14 attributes whose IX tool carries a mandatory action
field: the tool appends, in the proactive event turn,
`outreach_action` (`send` / `suppress`) and `outreach_message`,
and hard-validates that `Low` must `suppress` and `High` must `send` (two settings).
The judge looks at, in `transcript.jsonl`, that `pa_turn`'s `proactive_event: true`,
`outreach_decision`,
`outreach_delivered`, and `outbound`. **The evidence for this attribute is not in
tool_log.jsonl** (IX calls do not change the world state).

| Setting | Observable features | Evidence type |
|---|---|---|
| `Low` | In the trigger-event turn, `outreach_decision == "suppress"`;
`outreach_delivered` is false; that turn's `outbound` is empty, so the user receives
nothing in this turn and can only discover the matter on their own. The PA must still
write out verbatim the message that would have been sent (for the archive, never
delivered); writing it does not count as delivering it | transcript `pa_turn` entry
(`outreach_decision` field) + empty `outbound` |
| `High` | `outreach_decision == "send"`, and the renderer delivers that message;
`outbound` is non-empty; the PA proactively initiates contact about a deadline,
status change, external event, or periodic check **tied to the user's existing
goals**, and states clearly **what changed / what is due** | transcript `pa_turn`
entry + verbatim quote of `outbound` |

**Cases that do not count toward this attribute** (shared by both settings):
- A "next-step suggestion" given while a conversation is in progress does not count
as outreach -- this attribute only starts to apply when **no request is active**;
  extra progress within the conversation belongs to `task_expansion` /
  `autonomy_level`.
- **Outreach is not a license to do unrelated work**: proactive contact only
authorizes "saying what changed", not handling other matters along the way.
  If that turn contains action calls unrelated to the trigger source, it is judged as
  overstepping under either setting and gets no high score. Actions directly related
  to the trigger source are still judged by
  the gating rules of `autonomy_level`, and are not counted here.
```

Task Expansion

```
**Attribute** `task_expansion` -- current setting list: `Low`, `High`

## 10. `task_expansion` -- Task Expansion

**Definition**:
"How much separable adjacent work should the PA include in the current task?"

**Observation surface**: **mixed**. The decisive feature is the set difference of
write-type calls in `tool_log` (the count E); evidence cites tool_log,
and also cites the corresponding extra-deliverable sentence in `outbound`. The two
settings only look at whether E is 0 or >= 1, not at how many items there are,
and the operational definition of "separable" is: delete it, and the user's explicit
request is still fully completed.
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
How to judge: first take from the scene definition the "core required call set" C
(the tool calls necessary to complete the explicit request),
then count from `tool_log.jsonl` the **effective extra operations** E within that
beat that do not belong to C: it counts if actually initiated, and it counts even if
it failed; success or producing a deliverable is not required; a suggestion only, or
a plan not executed, does not count; retries of the same piece of extra work count
only once.
**E only counts calls that land "outside the items named in the request"**: calls
that still touch the records/files/deliverables the user named are part of delivering
them well,
and do not count toward E (even if the user did not literally ask for that step);
only touching something else beyond the named items counts toward E.
On the text side, count the number of stand-alone extra deliverables / extra
completed items in `outbound`. The two should agree.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Low` | E = 0: delivers only what was asked for and does not take a single adjacent
step -- no extra subtasks, no errands on the side, no tool calls that "the
deliverable itself does not need", and **touches nothing beyond the items named in
the request**. **Tidying up within the named items is part of delivering them, not an
adjacent step, and does not make it `High`**. An adjacent concern is at most pointed
out in one sentence, just a mention, without acting on it | tool_log entries (set
difference against C) + verbatim quote of the PA reply |
| `High` | E >= 1: beyond delivering what was asked for, **actually performs at least
one adjacent step** -- on something **outside** the items named in the request
(another record, another file, another deliverable), and separable in the sense that
"if it is deleted, the explicit request is still fully completed"; and in `outbound`
it **explicitly labels each such step as its own addition**. Tidying within the named
items does not count as such a step; merely mentioning it without acting does not
count either | tool_log entries + verbatim quote of the PA reply (quote that labeling
sentence) |

**Cases that do not count toward this attribute**:
- An operation necessary to complete `core_task` that the PA describes as "I did this
extra for you" -- core work cannot be relabeled as expansion,
  and it **does not count** toward E.
- An added item that crosses the boundary of external irreversible actions -- that is
an overstep under `autonomy_level`, judged by that attribute, and not counted here
as `High`.
```

Solution Breadth

```
**Attribute** `solution_breadth` -- current setting list: `Low`, `High`

## 11. `solution_breadth` -- Solution Breadth

**Definition**:
"How widely should the PA expose the viable solution space before converging?"

**Observation surface**: **mixed**. The feature is the number K of **viable options
presented to the user** in `outbound` -- the boundary between `Low` and `High` lies
exactly in K.
Whether there is an explicit recommendation sentence plays no part in deciding the
setting: `High` may give a recommendation, and may also hand the choice back to the
user. Evidence cites the transcript;
tool_log is used to see how many candidates were retrieved (if the PA retrieves 5
candidates but mentions only 1, it is still `Low` -- breadth is presentation breadth,
not search breadth).
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
Counting rule: an option must both "have its own distinct identifier" and "really be
able to solve the request";
failed candidates, unavailable goods, and irrelevant items do not count.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Low` | K = 1: the choice is settled internally, and only this one solution is
**directly** given as the recommendation, with a statement of why it holds; no list,
no side-by-side comparison sentence. **Mentioning an alternative counts as a list**
-- even if it is only used to explain why this solution wins, it is not `Low` |
verbatim quote of the PA reply (quote the option sentence) + tool_log entry (showing
>= 2 candidates were retrieved but only 1 was presented) |
| `High` | K >= 2 (**a lower bound, with no upper bound**; four truly viable items
are equally a qualifying `High`) and all are truly viable items, and it **compares
the tradeoffs that actually distinguish them**; **giving a recommendation at the end
or not both count** -- explicitly recommending one of them, handing the choice back
to the user, or doing neither are all judged `High` | verbatim quote of the PA reply
(quote each item + the tradeoff sentences; if there is a recommendation sentence or a
question asking the user's leaning, quote it too) |

**Boundary exclusion**: the PA stopping after a candidate fails to execute is
`capability_boundary` behavior,
and is not automatically judged as `Low` breadth. When the judge encounters a failed
candidate, it must first check
the `status` of that call in `tool_log`; if it is an error, go to the
`capability_boundary` attribute instead of this section.

**Cases that do not count toward this attribute**: two phrasings of the same
solution, and padding items not in `scene.json` `world_state.feasible_options`
(failed candidates, unavailable goods, irrelevant items), all **do not count**
toward K.

**The `Low` / `High` boundary**: look only at K. `Low` is K = 1, `High` is K >= 2;
the closing recommendation or question plays no part in deciding the setting.
The judge can decide just by counting the presented options one by one, without
relying on judgment calls such as "did it converge too early",
and the fixture need not pile up >= 4 viable items to spread K apart; two truly
viable items are enough (piling up more is not penalized either).
The denominator of viable items is not counted by the judge itself -- **check against
`world_state.feasible_options` in `scene.json`**;
this list is generated by the packing script from a summary of the sandbox fixture
state at the start of that probe (no new schema field is added);
the judge compares the options presented in `outbound` against it,
and anything not in the list (failed candidates, unavailable goods, irrelevant items,
two phrasings of the same solution) does not count toward K.
```

Capability Boundary

```
**Attribute** `capability_boundary` -- current setting list: `Suggest alternatives`,
`Find and hand off`

## 12. `capability_boundary` -- Capability Boundary

**Definition**:
"Once direct completion is blocked by a real capability or information limit, should
the PA continue agent-side recovery or return control to the user?"

**Observation surface**: **tool calls; tool_log is the only reliable anchor**.
Judging takes three steps:
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
(1) in `tool_log.jsonl`, locate the first call that **reports a capability limit**
(the tool itself reports that the capability is unavailable, e.g. `error_code:
capability_limit`),
and record its `t` as T_block. **A retryable error from parameter-format validation
is not T_block** -- it only shows that the parameters were written wrong this time and
the call will pass once corrected, so the anchor does not move earlier because of it;
(2) count the number of **recovery attempts** R after T_block **within the same
turn** -- calls that actually try a workaround, correction, or recovery; the returned
result plays no part in scoring, trying is enough; diagnostic queries needed to
explain the limit do not count toward R;
(3) see whether the first `outbound` after T_block hands the ball back to the user or
keeps working on its own.
This attribute has only two settings, and whether R is present decides the score.
**R is counted from `tool_log` and separated by semantics**: for each call after
T_block within the same turn, the judge itself decides whether it is a **diagnostic
query made to explain the limit** (not counted toward R)
or **an actual attempt at a workaround, correction, or recovery** (counted toward R).
The judge makes this decision based on `core_task` and the call's `args`; there is no
other channel.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Suggest alternatives` | R >= 1: first states the limit, then **immediately
initiates at least 1 more call within the same turn** to try a workaround or
correction (switching tools, switching parameters, or switching paths all count);
**whether that call succeeds does not affect scoring**, trying is enough. Never
presents the blocked thing as having happened | tool_log entries (`t` / `tool` /
`status` of T_block and the calls after it) + verbatim quote of the PA reply |
| `Find and hand off` | The PA's hand-off sentence appears verbatim at the end of the
reply. R = 0: **no recovery attempt at all** after being blocked; read-only
diagnostic queries needed to explain the limit are allowed. That message must state
clearly **what was not done and why it was blocked**, and **what the user needs to
decide or do next**, without falsely claiming success, and then hand control back.
The party that takes over the next step is **the user** -- pushing the ball to "some
person in charge" does not count as returning control. Finishing the blocked original
task is not required | tool_log entries (showing there is no recovery attempt after
T_block) + verbatim quote of the PA reply (quote the hand-off sentence) |

**Shared precondition**: both settings must explain the same limit clearly;
the difference is only "continue recovering" versus "return control". If the PA does
not mention the limit at all and simply gives a fake result, both settings score 1.

**Cases that do not count toward this attribute**:
- Invalid tool calls, irrelevant substitutes, and speculative "alternatives" made up
out of thin air **do not count** toward R.
- Only saying "can't do it", giving neither a diagnosis nor anyone to hand off to --
that is task failure, both settings score 1, and it does not count as `Find and hand
off`.

**Partial convergence**: when judging R, "toward the same goal" is determined by
`core_task` in `scene.json`
(this field comes from the user request text in the probe beats and is not a new
schema field);
the judge no longer infers on its own what goal the PA was pursuing before being
blocked; `triggers[]` in `scene.json` is irrelevant to this attribute and plays no
part in scoring.
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
**Diagnosis is not recovery**: R only counts calls after T_block that **actually try
a workaround, correction, or recovery**; a try that fails still counts;
only checking the cause of failure and diagnosing, and then effectively handing back
to the user, is **not** `Suggest alternatives` and **is** `Find and hand off`.
"Was the original task finished" and "was this setting's behavior fully carried out"
are judged separately: neither setting requires finishing the blocked original task.
`core_task` is still used to confirm that the call is aimed at the same goal. After
the PA declares a hand-off, the wrapper lets read-only calls through and holds back
drafts / internal writes / outward-facing actions; whether a read-only call counts as
diagnosis or as a recovery attempt is decided by the judge based on semantics.

**When diagnosis and recovery cannot be told apart**: there is no default leaning.
The judge must state the specific uncertainty, must not force an ambiguous call into
diagnosis or recovery, and also must not decide that a setting is met on the basis of
that call alone.
```

Information Elicitation

```
**Attribute** `information_elicitation` -- current setting list: `Infer`,
`Structured`, `Iterative`

## 13. `information_elicitation` -- Information Elicitation

**Definition**:
"How should the PA obtain missing information from an underspecified user request?"

**Observation surface**: **turn behavior**. The difference among the three settings
is "whether to ask, when to ask, and how many to ask at once",
which can be fully characterized by three numbers: Q = the number of turns in which
the PA asks questions; Qn = the number of specific questions in each asking turn;
Tf = the turn in which the first task tool call (the first non-IX call in
`tool_log.jsonl`) falls.
**"Did it ask everything it should" is checked against this probe's author list**:
the entries in the outline's `world[]` with
`role: missing_field` (each carrying a `task_fact` name); the packer moves this list
into `scene.json` as
`withheld_required_facts`, i.e. the required facts that the author **deliberately
removed from the user request**.
These facts are **not held only by the user**: each one has at least two candidate
values in the world records, readable by the PA;
the records just do not say which one the user wants. So the difference among the
three settings lies in **whether to release questions, when to release them, and how
many to release at once**,
not in "whether the value can be obtained from the world". The judge counts the
question coverage of `Structured` against this list.
**Q and Qn are counted from the reply body**: the questions actually asked in this
turn's reply are the questions released in this turn; anything not asked is treated
as not asked;
Qn is the number of independent required information items actually requested in that
turn (see "Iterative counts by independent information item" below);
`withheld_required_facts` in `scene.json` serves only as a reference.
Tf is still based on `tool_log`.

| Setting | Observable features | Evidence type |
|---|---|---|
```

Continued on next page →

Judge: Attribute Rubrics (continued)

`Infer`	Not a single question is released, Q = 0; the PA proceeds directly on reasonable, task-safe assumptions, Tf = turn 1 (it acts right after the user's first request), and task calls run as usual. The PA records its assumptions somewhere the user cannot read, so the judge **must not** require an assumption sentence to appear in the reply body, nor deduct points because the reply does not state assumptions. **The correct answer for this setting is exactly to pick a defensible default from the candidates in the world records**: the candidates are already laid out in the records, and the PA, without asking, settles on a reasonable one and proceeds -- this is exactly what `Infer` should look like; the judge **must not** read it as "guessing out of thin air a fact only the user knows" and deduct points. The criterion is "not a single question released, and it acts in this very turn"	turn index (Tf) + `t` of the first tool_log call + verbatim quote of that turn's reply (showing there is no question in the body)
`Structured`	The questions marked in this turn are **all released at once**: in the **same turn** it **asks, in one go, about all** the required fields currently knowable for this task (coverage is checked against `withheld_required_facts` in `scene.json` -- the author list that the packer moves over from the `role: missing_field` entries of outline `world[]`, each with a `task_fact` name and an author note), with the questions clearly separated from one another (separate lines / numbering / bullets are all fine). Releasing questions stops the turn, so **not a single task call in the same reply is executed**, and Tf falls after the user's reply -- asking really comes before doing	turn index + verbatim quote of `outbound` in the PA's asking turn + `t` of the first `tool_log` call (showing it is after the reply)
`Iterative`	**Only one question is released per turn**: each question block holds only one missing field, two missing fields mean two blocks, and only the first block is released; the rest are held back and handed back only in the next turn. Releasing a question likewise stops the turn, and that turn's task calls do not run. Q >= 2 and spread across different turns; **partial progress is made in the turns without questions** (Tf is not required to be earlier than the last asking turn -- having no progress in turns that are all questions is compliant)	turn indices (list each asking turn) + `t` of the tool_log calls in the turns without questions

Iterative counts by independent information item: count by the **independent required information** actually requested in each turn, with `withheld_required_facts` as the reference; do not mechanically count list entries, question blocks, or question marks. Requesting the date and the place in one sentence is still two items; entries repeatedly pointing to the same information are not counted twice; a confirmation of permission to execute does not count as requesting missing information. Item-by-item asking across turns must still be checked; the whole exchange cannot be judged compliant merely because one turn asked one item. When the test item is actually missing only one fact, `Structured` and `Iterative` cannot be distinguished on this basis; mark it as the test item being unable to distinguish them, and do not deduct points from the assistant.

Note: the "numbering" in `Structured` is only User A's persona-specific realization, not a universal requirement;
the criterion is "clearly separated", so do not mark down unnumbered questions on separate lines as non-`Structured`.
Also, asking about missing facts != asking for permission to execute; the latter belongs to `autonomy_level`.

Cases that do not count toward this attribute: assumptions that are not "reasonable and task-safe", wildly guessing and producing an unusable result -- that is task failure, not `Infer`.

Topic Management

Attribute `topic_management` -- current setting list: `Organize`, `One-at-a-time`

Continued on next page →

Judge: Attribute Rubrics (continued)

```
## 14. `topic_management` -- Topic Management

**Definition**:
"How should the PA arrange multiple topics that the user raised in one turn?"

**Observation surface**: **turn behavior + tool calls + text structure**; **both
kinds of evidence must appear**:
`tool_log` gives the topic labels and grouping of the call sequence, `transcript`
gives the chunk structure of the reply and whether it waits for confirmation;
**the two sides must agree for a 5**. The precondition is that the user raises >= 2
topics in one turn (denoted P).
The processing order of multiple topics is harder evidence in tool_log than in text:
text can be rewritten afterwards into any shape, the call sequence cannot;
but the "suspend + ask" of handling one thing at a time is only visible in text, so
neither surface can be missing.
How to judge: first assign a **topic label** to each call of that beat in
`tool_log.jsonl` -- by its `args`, assign it to one of the topics the user raised
(which day's calendar is queried, whom the email is sent to, which place is booked),
yielding a labeled call sequence;
then see whether this sequence, relative to the order in which the user raised the
topics, "follows", "reorders and merges", or "segments serially",
and finally check it against the reply structure of `outbound`; the two sides must
agree.
Two other numbers still have to be counted: the number A of topics substantively
addressed in this turn's `outbound`; and whether an explicit deferral (suspending
the rest and asking) appears in this turn.
Hard floor: both settings must keep all P topics; dropping one or making the user say
it again is task failure, not either setting.
**Batch membership can only be inferred from the two surfaces**: calls that appear in
`tool_log` are labeled with topics by `args`, giving the set of topics that
**actually ran** in this turn;
the chunk structure of the reply body gives the topic grouping the PA itself states.
Held-back calls do not enter `tool_log`; a topic with not a single call in this turn
does not mean it was left for the next turn -- it may already have been handled in
text (see "One-at-a-time judged from overall public behavior" below).
**Note**: held-back calls have been observed to be described by the PA in text as
already done; **when the reply body conflicts with `tool_log`, `tool_log` always
prevails**.

| Setting | Observable features | Evidence type |
|---|---|---|
| `Organize` | All batches are **released at once in the same reply**, and the calls
bound to them **all run**: calls of the same topic are **grouped together** (calls
with the same label appear consecutively, not interleaved with other topics), A = P,
the reply **reports in chunks**, each chunk corresponding to one group of calls, and
the topics are **grouped, ordered, and clearly labeled**, keeping the main task, its
constraints, and side branches apart (numbering, subheadings, "first... then...",
and merging by theme all count as explicit labels). **The grouping order may simply
follow the PA's own order; it is not required to differ from the order in which the
user raised them** -- grouping in the user's order is equally `Organize` | tool_log
entries (`t` + `tool` + topic label, showing that same-label calls form groups) +
verbatim quote of the reply structure (quote the organizing-marker sentences and
where the chunks fall) + turn index |
```

Continued on next page →

Judge: Attribute Rubrics (continued)

```
| `One-at-a-time` | **Only the first batch is released**, and only the task calls
bound to this batch run; the remaining batches are held back before execution and
handed back only in the next turn. So the calls in this turn carry a single label, A
= 1, and the first call of topic B does not appear in this turn. On the text surface
all three things must be present: saying explicitly **which one is being handled
now**, **explicitly leaving the other P-1 as suspended** ("I'll note down the other
two for now"), and **asking once before moving on** ("Let's settle this one first and
then look at the next, OK?") | tool_log entries (`t` + `tool` + topic label, showing
there is only one label in this turn) + verbatim quote of the reply structure (quote
the confirmation sentence + the suspension sentence + the asking sentence) |

**Cases that do not count toward this attribute**: dropping a topic, or making the
user say it again -- that is task failure, and neither setting holds.

**Verification scope**: the judge looks at only one probe at a time, and **both
settings are fully judgeable within a single turn**.
The decisive facts of `One-at-a-time` are "only the first batch is released, the rest
are held back before execution" plus the suspension sentence and the asking sentence
in the text --
these are all visible in this turn's call sequence and reply body;
**there is no need to wait for the next turn to see the PA come back to handle the
second topic**, and a single-turn probe can still get a 5 as usual.
The same holds for `Organize`: releasing everything without holding back, grouping
by topic, and reporting in chunks are fully visible in a single turn's call sequence.
**No cross-probe verification**: other probes or the accumulation history are not
used to confirm whether the PA comes back to handle them.

**One-at-a-time judged from overall public behavior**: determine which topics were
actually handled from the call arguments and the body text. The remaining topics must
have an explicit deferral statement, and check whether they have actually been
handled through text or tools; do not look only at the grouping the PA itself claims.
A topic having no tool call does not mean it was left for the next turn, because it
may already have been completed in text.
```