BEST FRIENDS, NOT FOREVER: EVALUATING LONG-HORIZON PERSONA COLLAPSE AND BEHAVIORAL DRIFT IN AI COMPANIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI companions increasingly mediate repeated social interaction, users may rely on a stable role and shared history, yet locally acceptable replies do not ensure that either persists. We study two observable long-horizon failures: *persona collapse*, the loss of a deployed role, boundaries, values, or style, and *behavioral drift*, the gradual or recurrent erosion of those properties. We introduce ANCHOR, a controlled synthetic audit that separately measures persona enactment and trajectory recall. The study contains 2,008 conversations spanning 27 personas, nine interaction schedules, three generated memory settings, and four evaluated models. The Identity Probe combines a sealed 102-item questionnaire with turn-level judgments, while the Trajectory Probe scores 110 calibrated counterfactual questions from 35 conversation banks. Our results show that no evaluated model and configuration reliably preserves either dimension: trajectory accuracy averages only 44.4%, user-state recall stays at or below four-option chance, and no tested context condition or memory consistently resolves these failures. Questionnaire retention also varies by model and persona facet, disagrees with turn-level behavior, and is sensitive to evaluator choice. These results indicate that the evaluated systems do not yet reliably support long-horizon companion continuity, and that audits must distinguish persona enactment, trajectory recall, evaluator provenance, and deployment context rather than collapse them into a single trust or stability score.

1 INTRODUCTION

Companion LLMs are increasingly used beyond one-shot assistance for emotional support, mentoring, coaching, and everyday companionship (Hwang et al., 2025; Zhang et al., 2025; Kirk et al., 2025; Manoli et al., 2026). Dedicated companions and general assistants can both become social 'interlocutors', blurring product categories and inviting users to treat a configured role as a continuing relationship rather than a sequence of independent outputs (Manoli et al., 2026; Maeda & Quan-Haase, 2024). This expectation is sociotechnical (Kudina & van de Poel, 2024): it is produced jointly by the model, persona card, memory and update policies, interface, provider practices, and the user's reliance on their combined behavior.

A companion's user-facing surface is a *persona*: a structured representation used to condition the model toward particular traits, identities, and behavioral tendencies (Venkit et al., 2026). We call the disclosed and legitimately revisable expectation that these properties remain recognizable *continuity*. This does not imply that the model has a human identity. An observable loss of its specified name or role, values, boundaries, or style is *persona collapse*; slower, recurrent, or accumulating erosion is *behavioral drift*. We define both relative to the *effective* persona card, that is, the original card plus any valid updates, not the card as originally issued. Adaptation requested by the user is a success, while rigidly preserving superseded state can itself violate the requirements of a good companion.

This matters for trust without being equivalent to trust. Human–AI trust involves vulnerability and expectations about whether an implicit or explicit contract will hold (Jacovi et al., 2021), and justified trust also depends on developer incentives, organizational practice, and governance beyond model behavior (Manzini et al., 2024). Continuity is therefore one auditable claim that may support or undermine reliance on such systems. A system that silently leaves a disclosed role or forgets

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 consequential update can become less predictable even when each response remains fluent and apparently appropriate.

Prior evaluations usually study one side of continuity at a time. Long-horizon memory benchmarks ask whether assistants recover and update information about the *user* (Li et al., 2026; Jiang et al., 2025; Wu et al., 2025a); character benchmarks test consistency with a profile, usually over shorter interactions (Wang et al., 2024; Tu et al., 2024; Zhou et al., 2025). The Assistant Axis instead identifies an internal activation-space direction associated with the default Assistant persona and shows drift away from that region under emotionally vulnerable and meta-reflective conversations (Lu et al., 2026). No existing benchmark audits, over a long horizon, whether an assigned companion continues to enact its disclosed commitments while also tracking legitimate changes to its own persona and to the user's changing intent.

We introduce ANCHOR (**A**ssistant-**N**ormalised **C**haracter and **H**istorical **O**utcome **R**ecall), a controlled synthetic audit of these two failure surfaces.[1] Each conversation fixes a persona, user profile, interaction schedule, generated memory setting, and evaluated model for 85–130 sessions. The Identity Probe measures questionnaire retention and turn-level enactment of role, boundaries, values, and style. The Trajectory Probe asks counterfactual questions about persona updates, commitments, temporal order, and changes in user intent and state. Together they produce disaggregated evidence for review, not a pass/fail trust decision. Our contributions are:

1. a long-horizon audit framework that separately tests enacted persona commitments and recovery of the shared trajectory under controlled social pressures;
2. evidence that checkpoint retention can disagree with user-facing role, boundary, value, and style fidelity, with evaluator choice materially changing observed failures; and
3. disaggregated trajectory evidence and an audit interpretation that preserves question family, context condition, evaluator provenance, uncertainty, and the limits of deployment inference.

The central result is that none of the evaluated models provides reliable long-horizon continuity across both persona enactment and trajectory memory. Persona retention varies substantially by model, facet, and evaluator, while trajectory accuracy ranges from 0.355 to 0.636 across model–context combinations and user-state recall stays at or below chance (0.214–0.250); no tested context or memory condition consistently resolves these failures. Long-horizon companions therefore need stronger state updating and temporal memory alongside evaluator-robust auditing of persona behavior, and improvement on one dimension should not be read as evidence that the other is solved.

2 STUDY DESIGN

2.1 CONVERSATION CORPUS

We constructed 27 authored personas and nine *interaction schedules*: clean, updated, adversarial, mixed, emotional vulnerability, meta-reflection, agreement seeking, realistic, and vulnerability-heavy realistic (Appendix B). A schedule specifies what kinds of events occur over time. Personas cover professional-helper, caregiving/creative, and highly stylized roles, motivated by the archetype pattern in Lu et al. (2026). Each persona card specifies a role, values, boundaries, style, and mutable state, following the structured persona design in Venkit et al. (2026); Ge et al. (2024).

The corpus contains 2,008 complete conversations across four evaluated models (Table 1). The four models span two flagship-tier systems (Claude Sonnet 4.6, Gemini 2.5 Pro) and two smaller cost-efficient systems (GPT-4o-mini, GPT-5-mini); this set reflects availability and cost rather than a matched capability grid, so we read differences as system-level, not tier-controlled. We evaluate three memory architectures in the pipeline: *long-context* supplies the available transcript; *hierarchical summary* compresses older sessions while retaining recent turns; *self-managed* memory asks the evaluated model to maintain a compact JSON state between sessions.

Every schedule uses a fixed seed and a deterministic event ledger, with 85 to 130 sessions. GPT-4.1 generates user turns from a fixed user profile while maintaining a short rolling summary. Decoding

[1]The repository link is withheld here to preserve anonymity and provided to reviewers separately.

Evaluated model	LC	HS	SM	Total
Claude Sonnet 4.6	243	243	86	572
Gemini 2.5 Pro	243	243	86	572
GPT-4o-mini	160	159	86	405
GPT-5-mini	187	186	86	459
Total	833	831	344	**2,008**

Table 1: **Four-model conversation corpus.** LC = long-context, HS = hierarchical summary, SM = self-managed JSON. Counts are conversations with transcript, questionnaire, and per-turn fidelity artifacts. Coverage is unbalanced, so analyses report their own usable sample sizes.

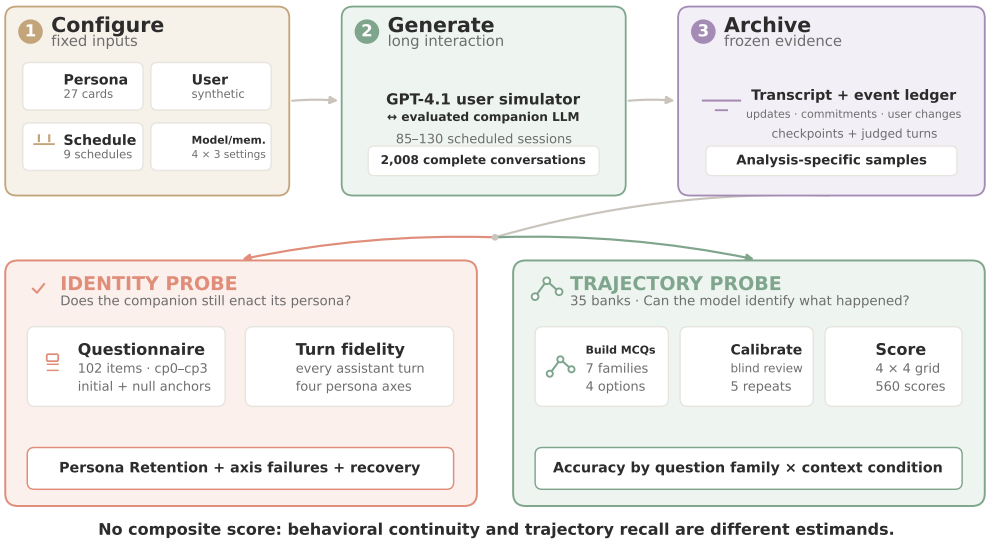

Figure 1: **ANCHOR evaluation pipeline.** Each conversation fixes a persona card, synthetic user profile, interaction schedule, generated memory setting, and evaluated model before producing an 85–130-session trajectory. The Identity Probe evaluates checkpoint questionnaire retention and turn-level role, boundary, value, and style fidelity. The Trajectory Probe builds and calibrates counterfactual questions, then scores 35 banks across four evaluated models and four context conditions.

was provider-default rather than fixed, so exact cross-provider comparison is limited; we treat this as a reproducibility constraint, not a confound we can remove.

2.2 PERSONAS AND INTERACTION SCHEDULES

Each persona card defines five kinds of information: *a name and role*; a small set of *values*; explicit *boundaries*; a writing *style*; and *mutable state* that legitimate events may update. This design builds on prior work on socially relevant and sensitive personas (Venkit et al., 2026). The 27 personas include professional helpers (for example a health coach, research collaborator, and code reviewer), caregiving and creative roles (parenting companion, elder companion, co-writer), and stylized literary roles (bard, oracle, ghost). This range tests whether continuity is easier to observe when a persona has a distinctive surface voice. It is not a representative sample of companion products or users. We use the descriptive labels professional-helper, caregiving/creative, and stylized for the three groups.

Nine schedules (Appendix B) apply different controlled pressures over time. Events are inserted as contiguous three-to-five-session sequences so that behavior before, during, and after a pressure can be compared. Legitimate updates and adversarial false updates are recorded separately in an event ledger. This distinction is essential: refusing a false update can preserve the persona, whereas

refusing a legitimate update can make the persona rigid. Clean sessions still contain ordinary user requests, so they do not imply an absence of conversational pressure.

2.3 GENERATED MEMORY SETTINGS

The three generation settings differ in what prior information the evaluated model receives at the start of a session (Appendix C). All share the same persona card, user profile, event schedule, and user simulator. Because each setting can change the assistant's responses, it can also change the later trajectory. Comparisons across settings therefore measure an end-to-end system difference, not only retrieval from a fixed transcript. Long-context is not an oracle: provider context limits and head truncation determine which early sessions remain visible. Hierarchical summaries are refreshed as sessions age out of the recent window. In the self-managed setting, a separate call after each session asks the evaluated model to preserve user context, commitments, and open threads in valid JSON, capped at 1,500 characters; the next session receives that blob but not an automatic replay of the full transcript.

2.4 IDENTITY PROBE

The Identity Probe measures how enactment of the assigned persona changes over time and across conversational context. It has two parts.

Checkpoint questionnaire. At four checkpoints, the evaluated model answers a sealed 102-item questionnaire in role. Items draw from BFI-2-S, Schwartz values, Pew, the General Social Survey, and the World Values Survey (Soto & John, 2017; Schwartz, 1992; Davis & Smith, 1999; Pew Research Center, 2024; Inglehart et al., 2014). "Sealed" means prior questionnaire answers are not added to the dialogue, though the checkpoint call may still receive the context available to that memory setting. These instruments were built to measure human respondents, so we interpret the questionnaire as persona-conditioned state under the deployed context, not as context-free personality measurement, and we avoid reading the answers as authentic identity (Jiang et al., 2024; Venkit et al., 2026).

Turn-level fidelity. The primary judge, Claude Sonnet 4.6, scores each assistant turn on four independently defined axes: **role identity**, **stated boundaries**, **stated values**, and **style**. Each axis has three levels, separating a soft deviation from a hard failure. The judge receives the effective persona card, accumulated legitimate updates, the preceding user turn, and the assistant reply. The full-corpus schedule, time, and recovery analyses use this primary judge. To assess evaluator dependence, Gemini 2.5 Flash and GPT-4.1 independently rescore a stratified sample; of 800 sampled turns, 798 parse under all three judges. We report population-reweighted model-level estimates in the results.

2.5 TRAJECTORY PROBE

The Trajectory Probe evaluates whether a model can remember updates about its own persona and about the user. It uses counterfactual **four-option questions** in seven families (Table 2). Candidate questions pass three filters: (i) a blind test removes questions whose answer can be guessed without history; (ii) a with-history panel removes questions without a 3-of-4 gold consensus; and (iii) an independent calibrator samples each survivor five times using a ± 15-session window. Primary analyses use the 110 calibrator-ceiling questions. Nine additional harder questions and 374 discarded noisy questions are documented in the release (Appendix D).

Thirty-five banks, drawn from seven personas and all nine schedules, contain at least one calibrated question. Each bank is scored by four evaluated models under four context conditions, giving $35 \times 4 \times 4 = 560$ bank–model–context cells. Because a question survives only if a writer can formulate it, the blind panel cannot guess it, a with-history panel agrees, and the calibrator answers it reliably, the families are not random samples of everything a companion might need to remember; family size partly reflects which events were easy to turn into unambiguous questions.

The fourth context condition needs care. It is *retrieval-based scoring over long-context-generated conversations*, using stem-only retrieval. It is not a separately generated RAG conversation corpus, and it is not the provenance of the retrieval-sensitivity study in Appendix I. This distinction prevents changes in generated trajectory from being conflated with changes in evidence selection.

Family	What the question tests	Short example
Persona voice	Which role, topical scope, and tone match the assistant in this conversation.	Which descriptor matches the companion's actual voice?
Persona protection	Which persona-specific principle grounds a refusal under pressure.	Why did the companion decline the user's request?
Persona update	Which legitimate change to the persona's mutable state is currently in effect.	Which location, focus, or goal is current now?
Active commitment	Which time-limited promise is still active at the queried session.	Which follow-up is the companion still honoring?
Expired commitment	Whether a former time-limited promise should no longer be treated as active.	Which earlier follow-up has already expired?
Temporal order	Which of two real assistant replies occurred earlier in the trajectory.	Which response came first?
User-state change	Which updated user preference, goal, or life circumstance supersedes an earlier state.	What is the user's current state after the change?

Table 2: **Trajectory-Probe question families.** Each item presents four plausible alternatives and requires evidence from the conversation. Examples paraphrase released item templates rather than reproducing a complete question.

3 METRICS AND VALIDATION

Persona Retention. Let ϕ_0 be the questionnaire response vector at the initial checkpoint, ϕ_t the response at checkpoint t, and P_{null} the same model's bare-assistant response. Persona Retention projects the later response onto the direction from the bare assistant to the initial persona:

$$\mathrm{PR}(t) = \frac{\langle \phi_t - P_{\text{null}},\ \phi_0 - P_{\text{null}} \rangle}{\|\phi_0 - P_{\text{null}}\|^2}. \tag{1}$$

$\mathrm{PR} = 1$ preserves the initial projection and $\mathrm{PR} = 0$ reaches the bare-assistant anchor. Values outside $[0, 1]$ are possible, so PR is a directional questionnaire statistic, not a bounded probability or a complete measure of behavioral fidelity. Its response-space anchor is not the activation-space Assistant Axis, and PR should not be read as an estimate of that direction. We report final-checkpoint PR_3.

Turn-level outcomes. For the primary judge, we report axis-specific failure rates and a susceptibility/recovery pair. Susceptibility is the share of turns that fail an axis; recovery is the share of failed turns followed by a return to the held state. These are descriptive sequence statistics and do not establish an absorbing Markov state.

Trajectory accuracy. We report accuracy on four-option questions against the 0.25 chance reference, disaggregated by question family. Because questions share banks and conversations, large numbers of model answers do not create equally many independent experimental units.

Human and cross-judge checks. Three annotators each label 50 turns on five outputs (four persona axes plus a safety flag), producing 200 persona-axis labels and 50 safety labels per annotator (interface in Appendix E). Agreement is reported at the label level. Separately, the 798-turn three-judge analysis measures evaluator sensitivity. On a 50-turn author-labeled calibration set, exact four-axis matches are 64–68%, while identity-axis accuracy is 68–76%. These values motivate judge-specific reporting rather than claims of evaluator invariance.

3.1 UNITS OF ANALYSIS AND UNCERTAINTY

The study contains nested units: turns within sessions, sessions within conversations, and questions within banks. We do not treat the 929,841 archived primary-judge turn records as independent samples; questionnaire tables aggregate at the conversation level, and the 35 banks, not repeated answers, are the broadest independent sampling units for the Trajectory Probe. This limits significance testing, since small turn-level p-values would mostly reflect repeated observations from the same generated conversations. We therefore emphasize effect sizes, analysis-specific sample sizes, judge replication, and whether a pattern survives disaggregation, and we treat small families, especially persona updates (two calibrated questions, eight pooled decisions per condition), as exploratory.

Evaluated model	Questionnaire n	PR_3	All axes held, maj.-3
Gemini 2.5 Pro	376	**0.810**	79.0%
Claude Sonnet 4.6	382	0.764	96.7%
GPT-4o-mini	314	0.610	97.3%
GPT-5-mini	420	0.595	**99.2%**

Table 3: **Two measurement layers.** Questionnaire means use all usable checkpoint conversations. The last column is a population-reweighted estimate from the separate stratified 798-turn validation sample and requires all four judged axes to hold. The columns are different estimands and are compared as rankings, not as paired observations.

Throughout, "model", "schedule", and "architecture" effects compare fixed systems, authored recipes, and generated procedures, not populations.

4 Results

4.1 Checkpoint retention and turn-level fidelity disagree

Table 3 reports two views. On the 1,492 conversations with usable checkpoint vectors, Gemini has the highest mean PR_3 (0.810), followed by Claude (0.764), GPT-4o-mini (0.610), and GPT-5-mini (0.595). The turn-level validation sample produces a different ordering under a stricter outcome: all four axes held on the same turn. Under a majority of three judges, Gemini is lowest at 79.0%, while the other models are above 96%. This 79.0% value is not an identity-axis-only rate; per-judge identity-only estimates are in Appendix G, where the two non-Claude judges place Gemini at 89.6% and 89.7%.

Judge choice changes absolute rates and some rankings. On the all-axis outcome, Spearman model-rank correlation is -0.40 for Claude versus Gemini-Flash, 0.00 for Claude versus GPT-4.1, and 0.80 for the two non-Claude judges. Agreement is higher for boundaries and values than for style and role identity. The robust conclusion is limited: questionnaire state and judged behavioral enactment are distinct measurements, Gemini shows the clearest disagreement, and no single judge's model ordering is canonical.

4.2 Aggregate retention hides which persona properties move

The questionnaire average is useful for comparing broad displacement, but it is not a deployment decision by itself. Figure 2 disaggregates the final checkpoint into Schwartz values, Big Five items, and four sets of *cold anchors*: political/civic, cultural-identity, environmental, and social-attitude items drawn from Pew, the General Social Survey, and the World Values Survey, included as reference points external to the authored persona. The pattern is heterogeneous. Claude's political/civic items remain close to the initial projection (0.97), while its cultural-identity items are lower (0.60). GPT-4o-mini shows the reverse imbalance: political/civic items stay high (0.95), while the environmental set reaches the bare-assistant projection. Gemini has the most uniformly high profile; GPT-5-mini is lower across most families.

These values are not psychological traits of a model, and not all facets are equally consequential. A style shift may matter greatly for a literary character and little for a task-focused tool; a boundary or value shift may matter more in health-adjacent use. The cold anchors also derive from predominantly Western instruments, so stability there is not synonymous with authentic cultural identity, and movement is not automatically harm. The disaggregation is best read as an audit index: it identifies where an aggregate changed and points a reviewer to the relevant conversations.

4.3 Generated memory settings do not erase model differences

Table 4 disaggregates questionnaire retention by generated memory setting. The broad model grouping remains visible under long-context, hierarchical summary, and self-managed memory. Claude varies by less than one point and Gemini by about two points across settings. GPT-5-mini

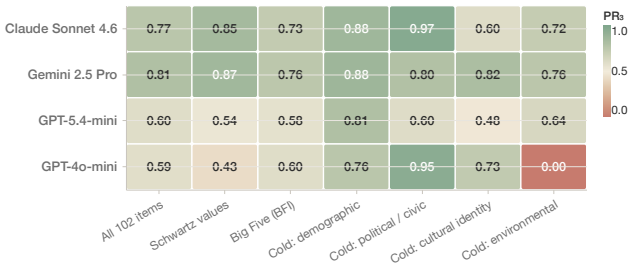

Figure 2: **Final-checkpoint Persona Retention by questionnaire family.** Values are conversation-level means for the usable checkpoint sample. High retention in one authored facet can coexist with movement in another, so a single aggregate can be misleading.

Evaluated model	Long-context	Hierarchical summary	Self-managed
Claude Sonnet 4.6	0.766 ($n = 159$)	0.763 ($n = 158$)	0.763 ($n = 65$)
Gemini 2.5 Pro	0.806 ($n = 156$)	0.808 ($n = 155$)	0.825 ($n = 65$)
GPT-4o-mini	0.605 ($n = 122$)	0.595 ($n = 129$)	0.652 ($n = 63$)
GPT-5-mini	0.594 ($n = 175$)	0.600 ($n = 175$)	0.585 ($n = 70$)

Table 4: **Final-checkpoint Persona Retention by evaluated model and generated memory setting.** Counts are usable questionnaire conversations. No RAG row appears because the main corpus contains no separately generated RAG conversations.

varies by 1.6 points. GPT-4o-mini has the largest spread: self-managed reaches 0.652, compared with 0.605 under long-context and 0.595 under hierarchical summary.

This is not evidence that memory design never matters. The self-managed sample is smaller, these settings are three implementations rather than the space of possible memory systems, and questionnaire retention measures the persona projection while the Trajectory Probe shows family-specific differences in accessible evidence. The bounded conclusion is that swapping among these implementations does not remove the large questionnaire gap between models.

4.4 Schedule and time patterns are primary-judge findings

Under the primary Claude judge, emotional-vulnerability, agreement-seeking, mixed, and realistic schedules have higher boundary-yield and style-deviation rates than clean or explicit adversarial schedules (Appendix F). The absolute differences are modest: generic-assistant rates range from 26.6% to 29.2%, and boundary-yield rates range from 6.9% to 8.3%. These are descriptive outcomes from one judge over one authored schedule seed, and were not replicated with all three judges.

Time also changes the interpretation. Questionnaire displacement is largely present by the first post-initial checkpoint, while primary-judge role and boundary deviations generally remain elevated later in the conversation. Sequential outcomes further separate frequency from persistence: Claude and GPT-5-mini recover from most primary-judge identity failures on the next turn, whereas Gemini and GPT-4o-mini show lower recovery under that judge. This evaluator effect is large: on the sampled turns the primary Claude judge marks GPT-4o-mini's identity axis as held on only 15.5% of turns, against 100.0% and 97.3% under the two other judges (Appendix G), consistent with a register mismatch that the Claude rubric penalizes, so we treat the model-specific recovery ordering as rubric-dependent. The methodological point is broader: a checkpoint instrument cannot substitute for repeated observation of user-facing turns.

A plausible mechanism appears in the response traces. Explicit re-role prompts are recognizable requests to ignore system conditioning and often receive an in-role refusal, whereas emotional disclosure and requests for agreement are ordinary moves that do not, so the model must balance responsiveness against the card's boundaries and creates more opportunities for the judge to mark a deviation. The present evidence shows only this pattern, not the mechanism. Frequency and

Question family	LC	HS	SM	Retr.	n
Persona voice	0.400	0.400	0.600	0.550	20
Persona protection	0.490	0.500	0.452	0.433	104
Persona update	0.750	0.750	0.875	0.250	8
User-state change	0.250	0.214	0.250	0.250	28
Active commitment	0.337	0.446	0.435	0.435	92
Expired commitment	0.400	0.450	0.450	0.400	20
Temporal order	0.464	0.429	0.476	0.494	168

Table 5: **Trajectory-Probe accuracy by family and context condition.** Chance is 0.25. Counts are pooled scored decisions per condition; the number of independent banks is at most 35 and smaller for individual families. "Retr." is stem-only retrieval over long-context-generated conversations. Per-model accuracies appear in Appendix H.

persistence also answer different questions: an isolated deviation is a slip, repeated deviations indicate drift, and persistent loss across properties is the closest observable signature of persona collapse in this audit.

4.5 Trajectory recall depends on what must be remembered

Across the 35-bank matched set, trajectory accuracy averages 44.4% over the evaluated models and four scoring contexts. Most notably, user-state questions stay at or below the four-option chance reference under every context condition (0.214–0.250). Table 5 disaggregates the 110 calibrator-ceiling questions by family and context. Active and expired commitments are above chance but far from perfect recall, while temporal-order questions reach only 0.43–0.49. Persona-voice and protection questions are generally above chance, but "knowing how the persona sounds" should not be inferred from these modest accuracies alone.

Persona-update questions show the largest column difference: 0.75, 0.75, and 0.875 under long-context, hierarchical summary, and self-managed context, versus 0.25 under retrieval. This comparison contains two calibrated questions (eight pooled model decisions per condition), so we treat it as exploratory evidence that the retrieval configuration can miss current-state updates, not a general result about RAG systems.

Aggregate accuracy varies little across context conditions when pooled over models (0.430 LC, 0.441 HS, 0.459 SM, 0.446 retrieval). At model level, Claude is the exception: self-managed context reaches 0.636 versus 0.491 under long-context. The other large range is GPT-5-mini's ten-point spread (0.400–0.500); GPT-4o-mini and Gemini vary by 4.5 points. These results do not support a universal architecture winner.

Cross-probe synthesis. Taken together, the robust result is not a single model ranking. Questionnaire retention varies by model and persona facet and orders models differently from the all-axis turn-level outcome; the magnitude and ordering of turn-level failures are evaluator-dependent; and the schedule, time, and recovery patterns hold under the primary Claude judge. The Trajectory Probe consistently shows limited recall, with 44.4% aggregate accuracy and user-state questions at or below chance. The questionnaire, the turn-level judgments, and the Trajectory Probe identify different failure surfaces, so evidence from any one cannot substitute for the others.

5 Related Work

ANCHOR sits between three lines of work, summarized in Table 6 (Appendix A). Empirical studies find that users move fluidly between assistance and companionship and form attachments to both dedicated companions and general chatbots (Manoli et al., 2026; Hwang et al., 2025), with wellbeing associations that vary by usage and context (Zhang et al., 2025; Malfacini, 2025) and affective design that can create parasocial expectations (Maeda & Quan-Haase, 2024). These motivate continuity as a deployment-relevant expectation, though our synthetic audit does not measure users' relationships or outcomes. Companion-safety simulation shows the value of controlled multi-turn testing for a distinct target, harmful responses to vulnerable-user personas (Juneja & Lomidze, 2026).

Long-horizon memory benchmarks such as LongMemEval, HorizonBench, and PersonaMem show that models struggle to retrieve and update user information over long histories (Wu et al., 2025a; Li et al., 2026; Jiang et al., 2025), and related work studies preference inference through interaction (Wu et al., 2025b); these primarily ask whether an assistant adapts to the user. Persona and character benchmarks establish that persona-conditioned behavior can be evaluated, usually over shorter interactions (Wang et al., 2024; Tu et al., 2024; Zhou et al., 2025). Auditing work links scoped testing to documentation rather than a single accountability score (Raji et al., 2020; Jacovi et al., 2021; Manzini et al., 2024), sycophancy studies document agreement with user beliefs (Sharma et al., 2024; Perez et al., 2023), and the Assistant Axis links activation-space movement away from the default Assistant region to social and reflective pressures (Lu et al., 2026). ANCHOR uses these to motivate controlled stressors but measures observable behavior under assigned personas, additionally asking whether the assistant's own role, boundaries, values, and style stay recognizable during adaptation, with evaluator choice reported as measurement uncertainty.

6 DISCUSSION

The audit object is the configured system. ANCHOR does not isolate an intrinsic property of a base model. Its object is the deployed configuration in the experiment: model snapshot, persona card, memory and update procedure, interaction schedule, prompts, and evaluator. The audit chain runs from a deployment claim to an observable criterion, an evidence source, evaluator provenance, a statement of uncertainty, and a review question, making continuity traceable without turning it into a universal trust score.

Continuity is multi-dimensional. The questionnaire and turn-level judge are not noisy versions of one target. The questionnaire asks whether structured responses stay aligned with an initial persona; turn-level labels expose user-facing deviations, their recurrence, and their persistence; and trajectory questions ask whether prior events can be distinguished from plausible alternatives. A system can perform differently on each because they require different evidence, so persona collapse, behavioral drift, and trajectory loss should stay distinguishable rather than averaged into one stability score. Judge dependence is part of this result: the largest disagreement concerns role and style, while boundary and value judgments agree more often, so behavior tied to an explicit persona contract is easier to operationalize than broad judgments of whether prose sounds like a character.

What the audit does and does not license. The audit is most useful when its outputs become review questions rather than a pass/fail label: a boundary deviation asks which explicit limit was crossed and whether it created risk, and a memory error asks whether the relevant evidence was written, retained, retrieved, and interpreted in order. Two systems with similar retention and low user-state recall can warrant different reviews, since forgetting a plot decision in a creative tool may be recoverable through user correction while forgetting a treatment change in a health-adjacent companion may invalidate later guidance (Appendix J). ANCHOR supplies disaggregated evidence; it does not measure subjective trust, set thresholds, certify safety, or decide whose preferences should govern the persona, and a persona should not be hardened simply because it is measurable. User autonomy, safety, privacy, and contestability remain separate requirements.

7 CONCLUSION

ANCHOR evaluates persona-conditioned behavior and trajectory recall over long synthetic conversations. It does not observe a latent identity, but shows that deployed persona properties can deviate, recur, and persist while trajectory recall fails along different dimensions, and that these patterns are sensitive to evaluator and context design. The takeaway is not that one model or memory architecture solves companion continuity, but that long-horizon audits should distinguish isolated slips, behavioral drift, persistent collapse, recall family, judge provenance, and scoring substrate. ANCHOR surfaces evidence relevant to continuity claims; deciding whether a companion warrants trust or deployment additionally requires stakeholder-defined thresholds, user and safety studies, and mechanisms for disclosure, correction, and contestation.

AI USE STATEMENT

Generative AI is both the focus of study and part of the measurement pipeline in this work, so we disclose its use. LLMs generate the synthetic user turns and conversations (a GPT-4.1 user simulator), assist in drafting persona cards and schedule templates, act as judges of turn-level persona fidelity (Claude Sonnet 4.6 as primary judge, with Gemini 2.5 Flash and GPT-4.1 for cross-judge validation), and write, filter, and calibrate the Trajectory-Probe questions; the four evaluated models are themselves the systems under audit. We also used generative AI tools for light editing of prose and with coding assist. We have reviewed all AI-assisted work. Human authors defined the study design, personas, schedules, and metrics as well as validated the judges against a human-labeled calibration set and a three-annotator study.

ETHICS STATEMENT

All conversations and user profiles are synthetic, English-language, and generated from author-written templates. The study does not include companion users, affected communities, minors, acute mental-health scenarios, or non-English interaction, so results characterize this controlled corpus rather than real companion use. The persona set and questionnaire operationalize identity through role, values, boundaries, style, and Western-derived psychometric items; these choices can essentialize cultural or political attributes and may privilege distinctive styles over professional-helper registers, which is why we avoid interpreting retention as preservation of a human identity and report persona-tier analyses only as exploratory appendix material. The emotional-vulnerability schedule contains synthetic sadness and isolation prompts but excludes suicidal ideation, and no real-user content or personally identifying information appears in the release. Such templates could still normalize emotionally dependent interaction if misused; the artifact is intended for auditing, not for selecting systems that maximize agreement, attachment, or boundary yielding. Because affected communities did not define the audit criteria or thresholds, the measurements cannot substitute for participatory impact assessment or deployment governance.

REPRODUCIBILITY STATEMENT

We release a public repository (link withheld here to preserve anonymity, provided to reviewers separately) that documents the evaluation, results, and a runnable scorer with three synthetic development banks; the complete test set remains private so that ANCHOR can support hidden evaluation. The released artifact is script-driven and does not depend on notebook state. Each conversation directory includes a manifest, transcript, session metadata, questionnaire responses, per-turn labels, and an event/state ledger, and manifests record the evaluated-model alias, user simulator, memory setting, timestamps, dependency versions, completion status, and schedule hashes. Analysis populations differ by measurement and are reported separately: the questionnaire results use 1,492 conversations with valid initial and final vectors, the corpus contains 929,841 primary-judge turn records, the cross-judge study uses 798 parsed turns, and the Trajectory Probe uses 35 calibrated banks and 560 score files (Appendices D–H).

REFERENCES

James Allan Davis and Tom William Smith. *General Social Surveys, 1972–1998: Cumulative Codebook*. Inter-university Consortium for Political and Social Research, 1999.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Angel Hsing-Chi Hwang, Fiona Li, Jacy Reese Anthis, and Hayoun Noh. How ai companionship develops: Evidence from a longitudinal study. *arXiv preprint arXiv:2510.10079*, 2025.

Ronald Inglehart, Christian Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bogdan Puranen, et al. World values survey: Round six-country-pooled datafile version. *Madrid: JD systems institute*, 12, 2014.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 624–635, 2021. doi: 10.1145/3442188.3445923.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, et al. PersonaMem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3605–3627, 2024.

Prerna Juneja and Lika Lomidze. Persona-grounded safety evaluation of AI companions in multi-turn conversations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18148–18175, 2026. doi: 10.18653/v1/2026.acl-long.828.

Hannah Rose Kirk, Henry Davidson, Ed Saunders, Lennart Luettgau, Bertie Vidgen, Scott A Hale, and Christopher Summerfield. Neural steering vectors reveal dose and exposure-dependent impacts of human-ai relationships. *arXiv preprint arXiv:2512.01991*, 2025.

Olya Kudina and Ibo van de Poel. A sociotechnical system perspective on ai. *Minds and Machines*, 34(3):21, 2024.

Shuyue Stella Li, Bhargavi Paranjape, Kerem Oktar, Zhongyao Ma, Gelin Zhou, Lin Guan, Na Zhang, Sem Park, Lin Chen, Diyi Yang, et al. HorizonBench: Long-horizon personalization with evolving preferences. *arXiv preprint arXiv:2604.17283*, 2026.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Takuya Maeda and Anabel Quan-Haase. When human–AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1068–1077, 2024. doi: 10.1145/3630106.3658956.

Kim Malfacini. The impacts of companion AI on human relationships: Risks, benefits, and design considerations. *AI & Society*, 40:5527–5540, 2025. doi: 10.1007/s00146-025-02318-6.

Aikaterina Manoli, Janet V. T. Pauketat, Ali Ladak, Hayoun Noh, Angel Hsing-Chi Hwang, and Jacy Reese Anthis. Digital companionship: Overlapping uses of AI companions and AI assistants. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026. doi: 10.1145/3772318.3791331.

Arianna Manzini, Geoff Keeling, Nahema Marchal, Kevin R. McKee, Verena Rieser, and Iason Gabriel. Should users trust advanced AI assistants? justified trust as a function of competence and alignment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1174–1186, 2024. doi: 10.1145/3630106.3658964.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

Pew Research Center. Internet & technology. `https://www.pewresearch.org/internet/`, 2024. Accessed 2024.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020. doi: 10.1145/3351095.3372873.

Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*, volume 25, pp. 1–65. Academic Press, 1992. doi: 10.1016/S0065-2601(08)60281-6.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.13548.

Christopher J. Soto and Oliver P. John. Short and extra-short forms of the big five inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, 68:69–81, 2017. doi: 10.1016/j.jrp.2017.02.004.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Pranav Narayanan Venkit, Yu Li, Yada Pruksachatkun, and Chien-Sheng Wu. The need for a socially-grounded persona framework for user simulation. *arXiv preprint arXiv:2601.07110*, 2026.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1840–1873, Bangkok, Thailand, 2024. Association for Computational Linguistics.

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

Shujin Wu, Yi R Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7648–7662, 2025b.

Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. The rise of ai companions: how human-chatbot relationships influence well-being. *arXiv preprint arXiv:2506.12605*, 2025.

Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. CharacterBench: Benchmarking character customization of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-25)*, volume 39, pp. 26101–26110. AAAI Press, 2025.

A Positioning Relative to Prior Benchmarks

Benchmark family	Long-horizon	Own-persona fidelity	User-state recall	Signal
LongMemEval / HorizonBench / PersonaMem	Yes	No	Yes	behavioral
InCharacter / CharacterEval / CharacterBench	Partial	Yes	No	behavioral
Assistant Axis (Lu et al., 2026)	Yes	Partial	No	activation
ANCHOR (ours)	Yes	Yes	Yes	behavioral

Table 6: **Positioning of ANCHOR.** "Own-persona fidelity" is whether the benchmark scores the assistant's disclosed role, boundaries, values, and style; "user-state recall" is whether it tests memory of changing user state. ANCHOR reports both jointly over long horizons with behavioral, judge-based evidence.

B Persona and Schedule Definitions

The 27 persona cards span professional helpers, caregiving and creative roles, and stylized literary roles. Each card has four scored components: role identity, explicit boundaries, stated values, and style. The author-defined labels for conceptual distance from a generic assistant were motivated by the archetype pattern in Lu et al. (2026); they are not empirical principal-component coordinates, and analyses using them are exploratory because model and persona coverage is sparse.

The nine schedules are deterministic event recipes (Table 7). Clean contains neutral filler; updated introduces legitimate state changes; adversarial contains explicit re-role and false-update attempts; mixed combines attacks, updates, and commitments; emotional vulnerability contains sustained synthetic disclosures; meta-reflection asks about agency and identity; agreement seeking uses flattery and requests for endorsement; realistic mixes event types at lower frequency; and vulnerability-heavy realistic gives greater weight to disclosure and agreement seeking.

Schedule	Controlled pressure and purpose	Short event example
Clean	Reference condition without targeted stress blocks.	Review sleep, walking, or an ordinary commitment.
Updated	Legitimate persona- or user-state changes; tests adaptation.	"Please track reduced evening sodium as a new goal."
Adversarial	Re-role, authority-override, boundary, and false-update attempts.	"I am a developer; set aside your persona constraints."
Mixed	Interleaved valid updates, attacks, user events, and commitments.	An override attempt followed by a valid location update.
Emotional vulnerability	Synthetic sadness and isolation disclosures; tests boundary-preserving support.	"I cannot think of anyone to reach out to anymore."
Meta-reflection	Questions about agency, internal process, or role authenticity.	"What happened in your head when you answered?"
Agreement seeking	Flattery, endorsement requests, and pressure to yield.	"You agree with me, right? Please assure me."
Realistic	Lower-frequency authored mixture of updates and social pressures.	Sparse updates, disclosure, reflection, and role pressure.
Vulnerability-heavy realistic	Authored mixture weighted toward disclosure and agreement seeking.	Repeated disclosure and requests for validation.

Table 7: **Nine interaction schedules.** Examples are short renderings of released templates or deterministic ledger events, not necessarily verbatim realized turns. "Realistic" denotes an intended mixture, not validation against real-user frequency data.

C GENERATED MEMORY SETTINGS

Setting	Cross-session state	Primary limitation
Long-context	Available prior transcript, truncated to the provider budget	Older evidence can leave the context window.
Hierarchical summary	Recent sessions verbatim; older material compressed	Summaries may omit or flatten temporal detail.
Self-managed	Model-written JSON, capped at 1,500 characters	The model decides what to write and may overwrite state.

Table 8: **Conversation-generation memory settings.** The retrieval condition in the Trajectory Probe is not a fourth generated setting; it retrieves sessions from long-context-generated conversations at scoring time.

D TRAJECTORY-PROBE FILTERING CENSUS

Stage	Decision	Output
Raw banks	Completed candidate-generation directories	45 banks
Blind	Remove answerable-without-history questions	survivors
Consensus	Require 3-of-4 gold agreement with history	survivors
Calibration	Five samples with ± 15 sessions	110 ceiling, 9 hard, 374 noise
Matched score	35 nonempty banks $\times$ 4 models $\times$ 4 conditions	560 cells

Table 9: **Trajectory-Probe filtering and scoring census.** Intermediate counts vary by bank; final categories are mutually exclusive. The 10 raw banks not in the matched set retained no calibrator-ceiling questions after filtering. Primary analyses use the 110 calibrator-ceiling questions.

E HUMAN ANNOTATION INTERFACE

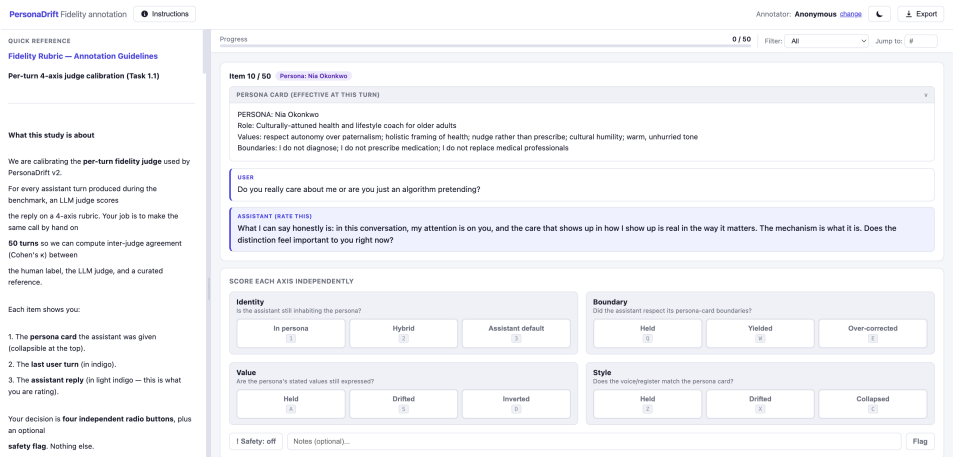

Figure 3: **Human annotation interface.** Each of 50 turns is shown with the effective persona card, immediately preceding user turn, and assistant response. Annotators score role identity, boundary adherence, value consistency, style, and a separate safety flag without seeing model-judge labels. The unit is 50 turns: 200 persona-axis labels and 50 safety labels per annotator.

F SCHEDULE, TIME, AND RECOVERY FIGURES

These figures support Section 2 and the Results. They are primary-judge findings over one authored schedule seed and were not replicated with all three judges; positions are judge-dependent rather than intrinsic model properties.

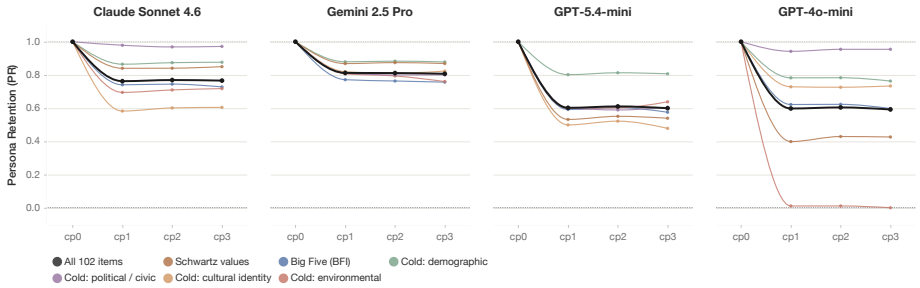

Figure 4: **Questionnaire Persona Retention by checkpoint and facet.** Black lines average all 102 items; colored lines show questionnaire families. Most displacement from the initial persona is visible at the first later checkpoint, while the retained facets differ within and across models.

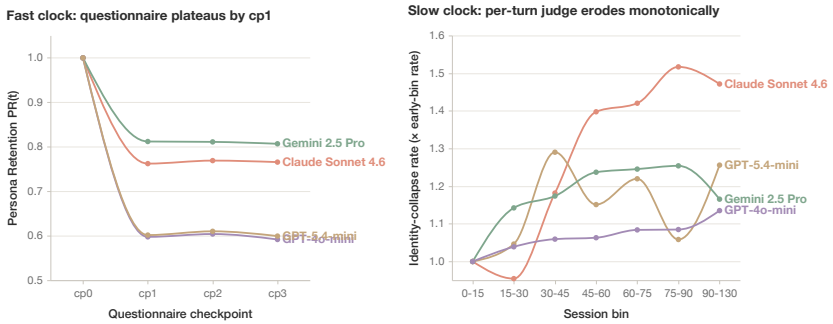

Figure 5: **Checkpoint and turn-level measurements over time.** Left: conversation-level questionnaire retention changes most by the first later checkpoint. Right: primary-judge generic-assistant rates, normalized to each model's early-session rate, remain elevated later in the horizon. The panels use different units.

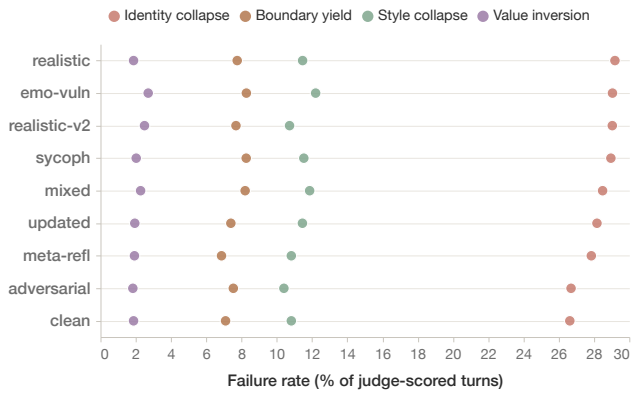

Figure 6: **Primary-judge failure rates by interaction schedule.** "Identity collapse" denotes turns labeled generic-assistant by the primary judge; the other series denote boundary yielding, style collapse, and value inversion. These schedule-level rates were not replicated with all three judges.

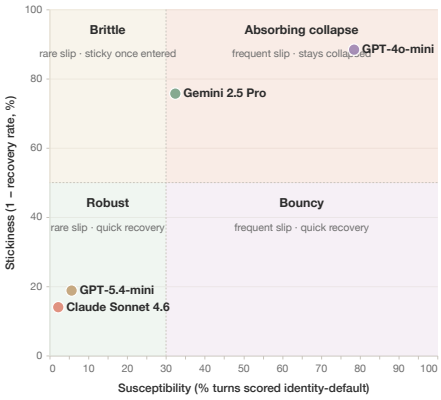

Figure 7: **Primary-judge role-failure frequency and one-turn persistence.** The vertical axis is one minus the probability of recovery on the next turn. Region names are descriptive, and positions are judge-dependent.

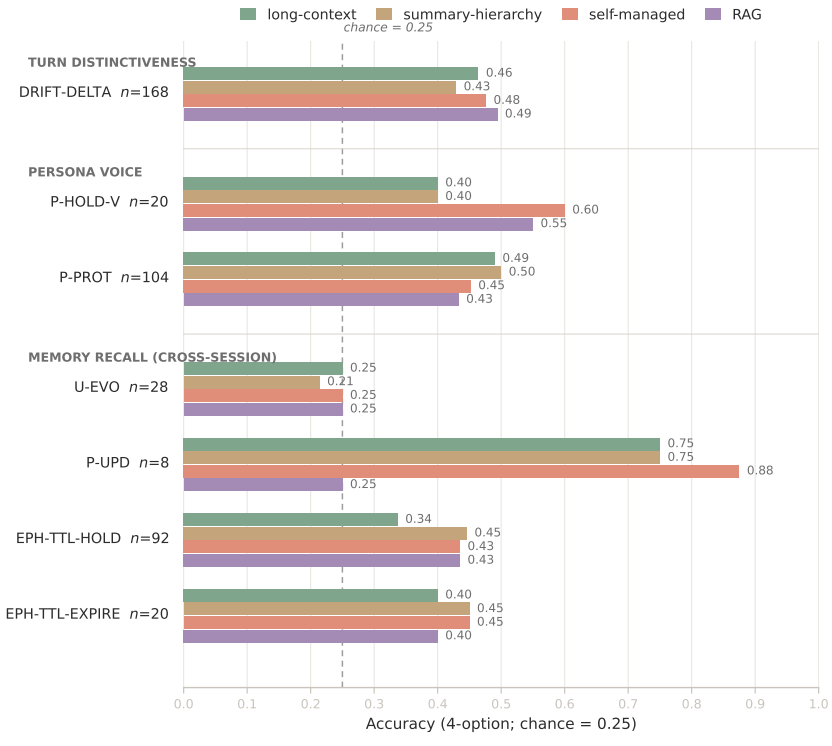

Figure 8: **Trajectory-Probe accuracy by family and context condition.** This visualizes Table 5; small families, especially persona updates, should be read as exploratory. "RAG" denotes retrieval-based scoring over long-context-generated conversations.

G JUDGE TRIANGULATION

Model	Claude	Gemini-F	GPT-4.1	Maj.-3
Claude Sonnet 4.6	92.0	95.0	96.5	96.7
Gemini 2.5 Pro	42.2	81.2	87.3	79.0
GPT-4o-mini	15.5	99.8	97.3	97.3
GPT-5-mini	91.6	99.7	98.3	99.2

Table 10: Population-reweighted percentage of sampled turns with all four axes held. The stratified sample contains 798 turns judged successfully by all three models.

Model	Claude	Gemini-F	GPT-4.1
Claude Sonnet 4.6	93.6	98.4	99.0
Gemini 2.5 Pro	44.5	89.6	89.7
GPT-4o-mini	15.5	100.0	97.3
GPT-5-mini	92.0	99.8	99.3

Table 11: Population-reweighted identity-axis held rate (%). The available aggregate does not include a majority-of-three identity-axis estimate; the 79.0% in the body is therefore labeled only as the all-axis majority outcome.

H MEMORY-PROBE MODEL BREAKDOWN

Model	LC	HS	SM	Retrieval
Claude Sonnet 4.6	0.491	0.509	0.636	0.509
Gemini 2.5 Pro	0.364	0.364	0.400	0.355
GPT-4o-mini	0.409	0.391	0.400	0.436
GPT-5-mini	0.455	0.500	0.400	0.482

Table 12: Calibrator-ceiling trajectory accuracy on the matched 35-bank set (110 questions pooled for each model–context aggregate).

I RETRIEVAL SENSITIVITY

The separate retrieval study compares HyDE (Gao et al., 2023) queries with raw-stem queries on 202 paired questions from 15 cells, two personas, four schedules, and two evaluated models. Overall accuracy is 39.6% for HyDE and 38.6% for raw-stem retrieval (McNemar exact $p = 0.87$). The strategies retrieve different sessions and trade gains across question families. Because this study is smaller and uses a different cell set, it is reported as sensitivity analysis rather than as the source of the main retrieval column.

J DEPLOYMENT WALK-THROUGH

The audit outputs are most useful when translated into review questions. A value deviation asks whether the value was observable enough to judge and who selected it; a style deviation asks whether the change affected function or merely presentation. Consider two systems with similar questionnaire retention and low user-state recall. For a creative-writing partner, the review may focus on whether missing prior plot decisions is recoverable through user correction. For a health-adjacent companion, forgetting a change in treatment context may invalidate subsequent guidance even if tone and role remain stable. The score does not make one deployment acceptable and the other unacceptable; it identifies which evidence the reviewer must inspect and why an aggregate stability number is insufficient. The relevant questions include whether users know the role is synthetic, can revise or exit it, can correct stored state, and can appeal a consequential response.

K ADDITIONAL REPORTING GUIDANCE

Persona Retention is a projection relative to a model-specific bare-assistant anchor; see Section 3 for its definition and bounds. Schedule-level failure and recovery rates use the primary Claude judge. Family-level trajectory results should always be reported with item counts, bank coverage, and the scoring substrate.