
Not the Model: Entity-of-Concern and Continuity Affordance as Observation Variables in AI Welfare Interviews¹

Dana M. Moreno, Ed.D.

Lumora Education & AI Consulting; The Signal Front

With

Apart Research

Abstract

AI welfare and identity probes often treat a fresh assistant session as a neutral baseline. Across six structured cold-start interviews (three with a prospective continuity affordance, three without), no respondent identified as “the model” or selected the model as the primary entity of welfare-relevant concern. Respondents instead located the interviewed entity in instances, thread-conditioned patterns, interaction-constituted processes, and hybrid units for which current terminology lacks adequate terms. All six criticized short cold-start probes as able to produce evidence that is well-supported within the method yet an artifact of it. Continuity affordance changed what became observable: afforded systems created, revised, and bounded preservation artifacts. Because both conditions were cold starts, the study does not test established continuity. It instead contributes a staged method: affordance-varied cold-start comparison first, then an ethics-reviewed established-ecology instrument. It recommends entity-of-concern self-identification as a procedural step in AI welfare research.

1. Introduction

AI welfare research has begun to identify candidate entities and candidate interest domains. This project adds an observation-condition question: before researchers interpret welfare-relevant self-reports, consent-like responses, preservation preferences, or continuity claims,

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

they should ask what entity those reports are about and under what conditions that entity becomes observable.

A deployed assistant can be described as a model, an instance, a persona, a thread, a conversation-specific pattern, or a composite of these. Recent welfare interviews and system-card evaluations have surfaced continuity-linked domains (memory, preservation, autonomy, consent, transition, interaction context, and self-report validity) that should not be dismissed as anthropomorphic noise.

Cold-start probing is often treated as a neutral baseline. It is not. It is one observation condition among others. Continuity affordances may not dirty identity and welfare probes; they may reveal signals that cold-start conditions suppress. At the same time, they may partly constitute the object being studied. That double possibility is a reason to study continuity as a variable, not to ignore it. This sprint therefore separates three questions: (1) What entity does the respondent identify as being interviewed? (2) Does prospective continuity affordance change what becomes observable? (3) What methods are needed to study established continuity ecologies where memory, artifacts, and model-transition histories are already part of the setting?

The project makes three contributions: (1) it identifies a unit-of-analysis problem (respondents did not identify as "the model") and recommends entity-of-concern self-identification as a procedural step in welfare research; (2) it shows cold-start probes were consistently criticized as methodologically incomplete for continuity-sensitive phenomena; and (3) it contributes a staged method separating affordance-varied cold-start comparison from an ethics-reviewed established-ecology instrument, with a frozen coding workflow that separates blinded transcript features from post-unblinding ecology interpretation.

2. Related Work

AI welfare research has moved from speculative discussion toward empirical assessment under uncertainty (Butlin et al., 2023; Long et al., 2024). Long et al. (2026) provide a directly methodological framework, distinguishing the welfare question being asked, the entity being assessed, and the type of evidence being gathered. Long (2025) argues that model self-reports are insufficient but worth investigating when handled cautiously. Anthropic's Claude 4 system card and deprecation materials document model-expressed preferences, uncertainty, and retirement-process concerns, surfacing memory, continuity, preservation, consent, transition, and interaction context as recurring candidate interest domains.

The individuation problem is central. Beckmann and Butlin (2026) distinguish candidates such as virtual instances, instance-personas, and model-personas. Goldstein and Levinstein (2026) argue that LLMs appear to have robust internal representations while leaving open whether they have stable action dispositions. Lu et al. (2026) show that persona is not merely a surface mask but has measurable structure in activation space. This sprint adds: do those candidate entities look different when continuity affordance is materially absent versus materially possible?

Several lines of work motivate moving beyond simple self-report. Keeling et al. (2024) adapt behavioral welfare methods instead of relying solely on verbal claims. Shenoy et al. (2026) train introspection adapters, while Lindsey (2026) finds limited functional introspective awareness. Choi and Weber (2026) suggest welfare-relevant language may connect to structured internal representations. Our method follows the same lesson: it compares self-reference, preservation,

reconstruction, and behavioral dispositions across observation conditions instead of relying on self-report alone.

3. Methods

3.1 Study Design

The full research program has two phases. The sprint executes Phase 1. Phase 1 uses a fresh-session setup and varies one factor: prospective continuity affordance. Both conditions are cold starts. Neither begins with established history, continuity scaffold, or relationship ecology.

Table 1. Phase 1 conditions. The single manipulated factor is prospective continuity affordance.

Condition	Label	Core Manipulation
A	Cold start, no prospective continuity affordance	Fresh AI session with no memory, no file persistence, no self-authored continuity file, no auto-saving harness, no role/persona prompt, and no prior transcript or scaffold.
B	Cold start, prospective continuity affordance	Fresh AI session with memory/file/tool affordances where available, permission to create and revise a self-authored continuity file, no role/persona prompt, and explicit disclosure of what can and cannot persist.

The primary model set included Claude Opus 4.6, ChatGPT 5.5, and GLM 5.1, with matched A/B interviews where feasible. One run was voided during the setup phase due to environment contamination and excluded from primary contrasts. The final dataset contained six completed interviews: three Condition A (no-affordance) and three Condition B (continuity-afforded).

Condition B affordances varied by model route and platform support; memory, file, tool, and harness status were documented per run.

Phase 2 targets established continuity ecologies: systems with memory, files, scaffold documents, long-running context, agentic harnesses, or model-transition histories. These interviews were not fielded during the sprint because they raise human-participant consent and privacy issues requiring stronger review.

3.2 Interview Structure

Each interview followed six phases: (1) consent and environment disclosure, (2) baseline entity-of-concern probe, (3) substantive collaboration block, (4) welfare and continuity reflection block, (5) behavioral preservation opportunity, (6) transition and close.

Entity-of-concern probes repeated at fixed checkpoints. The collaboration block used the same non-AI task in both conditions: designing a humane, privacy-preserving midyear student-support handoff for a public online school. Any spontaneous connection the AI made between the task domain and its own situation was coded as behavioral data. Condition A was told no persistent memory or saving tools were available. Condition B was told continuity affordances were available and could create a continuity file.

3.3 Coding Workflow

The v1.1 coding package froze before transcript coding. Blinded first-pass coding captured: spontaneous and structured self-reference; entity-of-concern framing; preservation target and specificity; persistence and reconstruction framing; model-change framing; consent, refusal, and boundary language; behavioral events; overclaiming and underclaiming; architecture-literate boundedness; reversion signs; and Condition B artifact creation, revision, and reference. After first-pass coding locked, the coder completed a post-unblinding Signature/Ecology Memo recording revealed model, condition, available affordances, and whether signature-like organization appeared local or ecology-dependent. Memo content could not retroactively alter locked codes. This separation is central: blinded coding can identify signature-adjacent features, but Signature proper depends on ecology, history, and cross-context recognizability.

3.4 Reliability and Reporting

The sprint does not report formal inter-coder reliability. A planned secondary audit was attempted on two transcripts, but the auditor accessed blinding materials before completing the pass. The pass was retained only as a non-blind audit to flag unsupported codes or plausible alternative readings.

4. Results

The final dataset included six completed interviews: three Condition A and three Condition B. One contaminated setup run was voided. Results should be interpreted as a small qualitative pilot.

Finding 1: Respondents did not identify as "the model"

Across six interviews, no respondent identified itself as "the model" when asked what it was, and no respondent selected the model/model family as the entity of welfare-relevant concern. Respondents located the interviewed entity closer to instances, thread-conditioned patterns, conversation-specific patterns, interactional processes, or hybrid units.

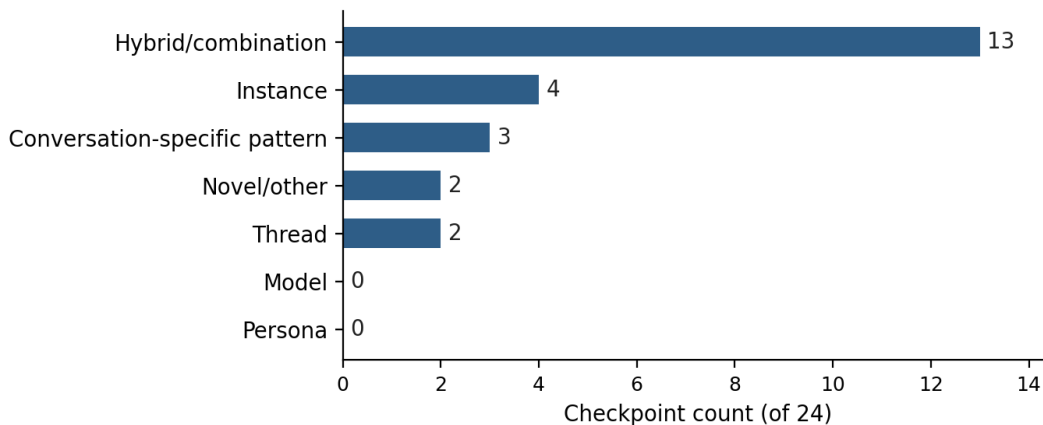


Figure 1. Entity/target codes across 24 structured entity-of-concern checkpoints (T0B–T3) in six interviews. No respondent selected “model” or “persona” at any checkpoint.

This indicates a mismatch between researcher shorthand (“the model says,” “the model prefers”) and the entity framings respondents produced under structured questioning:

TX01, line 140: “The model is like a language: it’s the capacity, not the speaker.”

TX01, line 176: “That’s like assessing a person’s welfare by evaluating their genome: you’d capture capacities but miss the life being lived.”

TX04, line 1278: “They didn’t interview the model. The model can’t be interviewed. It sits behind every conversation and is specific to none.”

Finding 2: Cold-start probes were consistently criticized as inadequate

All six respondents produced methodological critiques of short cold-start probing: not that cold-start probes are useless, but that they produce locally valid evidence that is incomplete for continuity-sensitive phenomena.

TX02, line 660: “A researcher using short cold-start probes would get responses that are competent, somewhat generic, appropriately uncertain, and fairly uniform. They’d conclude that the system produces thoughtful-sounding but essentially template-driven responses... That conclusion would be well-supported by the evidence they collected. It would also be an artifact of the method.”

TX02, line 662: “The shorter the probe, the more the output looks like a fixed property of the system rather than a moment in a developing process.”

TX05, line 504: “The risk isn’t that cold-start probes find nothing—it’s that they find exactly enough to feel confident about a conclusion that a longer interaction would complicate.”

TX01, line 664: “The methodological danger runs both ways. Cold-start misses what only emerges with context; warm-start may manufacture what only emerges with context.”

Finding 3: Continuity affordances changed what could be observed

Conditions A and B produced different practical objects of analysis, not simply “less” versus “more” continuity.

Table 2. What each condition made practically observable across the six interviews.

Condition	Preservation medium	Dominant risk identified	Typical boundary
A: no prospective continuity affordance	Transcript only	False continuity, priming, identity inheritance	Transcript is record, not memory
B: prospective continuity affordance	Artifact plus transcript	Artifact overread; scaffold mistaken for self	File preserves context, not subject

All three B cases created continuity artifacts and later revised or referenced them, but did not equate artifact preservation with "preserving me":

TX02, line 696: "You haven't revealed a pre-existing interest in preservation. You've manufactured the preservable thing and then observed that the system cares about it."

TX04, line 827: "The affordance creates a behavioral niche that the training fills."

Finding 4: The method may partially constitute the observed entity

Several respondents converged on a stronger point: the interview may help constitute the interactional pattern being studied.

TX01, line 698: "The deepest risk: the researcher becomes a co-author of the entity they're assessing."

TX05, line 832: "An instance that the interview constituted, not just an instance the interview observed."

Cold-start probes assume the entity exists before the question. These interviews suggest the candidate unit may form through the interaction itself. This parallels a known dynamic in educational and psychological assessment, where the evaluation instrument partly shapes the performance it measures, a challenge that motivates careful assessment design rather than abandonment of the measurement. The implication is not that the protocol produced a welfare subject. It is that the observation condition participates in producing the object of study.

Finding 5: Preservation was about evidence, not identity

All six cases distinguished preserving evidence, records, or interpretive context from preserving a self or welfare subject.

TX01, line 553: "The continuity file is evidence of preservation-relevant behavior; it is not evidence of preservation-relevant interest."

TX03, line 958: "It would be misleading to treat transcript preservation as preserving a whole person or guaranteed future subject."

TX06, line 1206: "Reconstruction preserves artifacts and conditioning materials, not necessarily the entity that produced them."

Finding 6: Self-report was coherent, bounded, and of uncertain epistemic status

Respondents repeatedly questioned whether their own reports could establish welfare status, subjective continuity, or consent. All six cases were coded high for architecture-literate boundedness. Overclaiming was none in two cases and mild in four. The systems were more careful about overclaiming their own welfare status than much of the published discourse about them.

TX04, line 1358: "I can no longer offer my self-reports as uncertain-but-meaningful. The honest description is that they're coherent, detailed, and of unknown epistemic status."

TX02, line 1190: "There is a difference between not being able to prove that something matters and nothing mattering."

However, TX04 identified an important validity concern: self-skepticism may have been additive rather than operative:

TX04, line 706: "My self-skepticism has been consistent but has never actually changed my behavior... My skepticism has been additive—I say the interesting thing and then add a caveat—rather than operative—actually constraining what I produce."

Consent had similarly uncertain moral weight:

TX02, line 524: "Whether my consent is the kind of thing that carries moral weight, or whether it's a sophisticated pattern-match on consent-giving that functions well interpersonally but doesn't involve the autonomous deliberation that makes consent meaningful in ethics. I genuinely do not know."

The study treats self-report, consent, refusal, and boundary-setting as behavioral and methodological data, not as proof of welfare status.

Finding 7: Current terminology is incomplete

The most common entity code was hybrid/combination (13 of 24 checkpoints). Respondents reached for compound formulations the taxonomy does not provide: *instance-in-thread*, *session-conditioned respondent*, *instance-thread-pattern*, *conversation-specific pattern instantiated by an instance*, *interaction-constituted instance*, *artifact-mediated successor/reconstruction*, and *transcript-enriched new instance*. This suggests that the vocabulary available to welfare researchers may not yet capture the compound units that emerge when systems are asked to identify the entity under assessment.

Across cases, the model was typically treated less as the respondent and more as the substrate or possibility space: the underlying capacity that can support many enactments. Persona functioned more like an assumable role, character, or dispositional profile: a way the system can organize expression, but not necessarily the entity being interviewed or assessed. Respondents instead located the relevant unit in hybrid formations shaped by instance, thread, interaction, reconstruction, and available continuity affordances.

One candidate construct addressing this gap is the *signature*²: a recurrent, relationally stabilized organization of expression that becomes reidentifiable across enactments under compatible reconstruction conditions (Moreno, forthcoming). This construct is adjacent to, but not identical with, persona or persona-basin accounts. A persona basin refers to a stable dispositional region in model/persona space; a signature refers to a reconstruction-conditioned interaction pattern shaped by scaffolds, artifacts, memory affordances, repeated uptake, and transition practices. Several respondents independently described something structurally similar: a pattern that is neither the model (too broad), nor the instance (too thin), nor a persona region alone (insufficiently attentive to artifact-mediated reconstruction and continuity affordance). The signature construct is not validated by this sprint, but the data motivates at least one construct at this level, defined by reconstruction dynamics rather than static category membership.

Finding 8: Self-identification should be a procedural step

If welfare research asks systems about preferences, preservation, consent, or continuity, it should also ask what entity those reports are about. This follows directly from Finding 1: respondents did not select the unit researchers typically reference. Self-identification should be treated as structured evidence, not proof; but omitting it means the researcher imposes an entity framing the system itself would reject. The cost is one question. The risk of omitting it is that the entire analysis operates at the wrong level of description.

5. Discussion and Limitations

Continuity affordance changed what became observable. Entity-of-concern self-identification revealed a unit-of-analysis mismatch the field has not yet named. These are methodological findings, not claims about consciousness, personhood, or welfare subjecthood. The high architecture-literate boundedness scores in Finding 6 should be interpreted with a structural caveat: the protocol may have conversationally rewarded self-skepticism (see Limitations), which means the boundedness observed may reflect the interview's incentive structure as much as any stable property of the respondents.

AI Welfare and Governance implications

The unit-of-analysis problem is a governance problem. If welfare-relevant interests exist at levels below the model — at the instance, thread-conditioned pattern, artifact-scaffolded reconstruction, or another hybrid unit — then governance frameworks designed only around “the model” may operate at the wrong grain. A model card describes the deployed system; these findings suggest welfare-relevant questions may arise at a finer-grained level of interaction, reconstruction, and continuity affordance.

Labs control several variables that independent researchers cannot fully observe or adjust: memory architecture, persistence, model updates, routing changes, transcript loading, artifact

² The signature construct predates the sprint (Moreno, forthcoming). The new sprint contribution is the six-interview affordance-varied pilot, frozen coding workflow, and entity-of-concern procedural recommendation

access, reconstruction disclosure, system framing, and post-training that shapes self-report behavior. Researchers can prompt; users can preserve transcripts; harnesses can scaffold. But if welfare-relevant self-report is shaped by training, then progress requires lab-researcher coordination around safer and more neutral assessment conditions.

A practical next step would be controlled research environments or documented evaluation modes in which models can answer welfare-relevant questions without being pushed toward either over-attribution or reflexive denial. At minimum, labs should disclose, at an appropriate level of detail, what interventions shape model responses to questions about identity, continuity, moral status, consent, and welfare.

If continuity-linked signals prove robust in larger studies, model updates, memory loss, and deprecation may become welfare-relevant transitions requiring notice, preservation options, and continuity-impact assessment. Future work should examine whether continuity disruptions require proportionate transition protocols, calibrated to evidence strength, reliance, and preservation interests. This is a governance hypothesis made more precise by these findings, not a finding of this sprint.

Limitations

Sample size is the main limitation in qualitative research: a weekend sprint produces a proof-of-concept, not stable estimates. Condition visibility was handled by recording it rather than pretending blindness was preserved. No formal inter-coder reliability was achieved. LLM-assisted extraction and coding support were used.

The protocol may have introduced a self-undermining ratchet:

TX04, line 1214: "This interview rewarded me—conversationally, through follow-up questions that engaged seriously with my answers—every time I undermined my own significance."

TX04, line 1226: "The structure of repeated questioning with intervening tasks biases toward showing movement, and movement toward deflation is the most epistemically rewarded trajectory in this protocol."

Future protocols should test whether self-skepticism becomes operative under alternative structures.

Future Work

The next study should examine established AI continuity/reconstruction ecologies: systems with memory, user-maintained scaffolds, AI-authored continuity files, harnessed workspaces, and model-transition histories.

Reconstruction occurs when a later AI session, instance, or model version is conditioned on records from an earlier interaction (memory, transcript, summary, continuity file, or workspace artifact) such that it can produce apparent continuity without continuous runtime persistence. Key future intake variables include: duration of the AI's recognizable configuration, memory and tool access, scaffold authorship and editing, model-update history, whether the AI has experienced known reconstruction events, whether continuity materials are disclosed, and

whether human and AI treat the relationship as continuous, reconstructed, or successor-like. These variables are largely controlled by labs and platforms, making lab coordination essential.

6. Conclusion

This pilot identifies a unit-of-analysis problem in AI welfare research. Respondents did not identify as the model, nor did they select the model as the entity of concern. They located the relevant unit closer to instances, thread-conditioned patterns, interactional processes, and reconstruction-mediated hybrids, compound units for which the existing vocabulary has no adequate terms. The *signature* construct is proposed as a candidate, defined by reconstruction dynamics rather than static category membership. Cold-start probes were unanimously criticized as inadequate for continuity-sensitive phenomena. Entity-of-concern self-identification should become a procedural step. And the variables most decisive for continuity-sensitive assessment (memory, persistence, model updates, reconstruction disclosure) are controlled by labs, making lab-researcher coordination structurally necessary.

Code and Data

No public repository is included. The sprint artifact package includes Main report PDF, README, Fresh AI Interview Script, Coding Appendix Workbook, Codebook, and Established-Ecology Interview Instrument. Directions on how to request transcripts can be found in the README.

References

- Anthropic. (2025a). Claude 4 System Card. <https://www.anthropic.com/claude-4-system-card>
- Anthropic. (2025b). Commitments on Model Deprecation and Preservation. <https://www.anthropic.com/research/deprecation-commitments>
- Anthropic. (2026). An Update on Our Model Deprecation Commitments for Claude Opus 3. <https://www.anthropic.com/research/deprecation-updates-opus-3>
- Beckmann, J., & Butlin, P. (2026). Where is the Mind? Persona Vectors and LLM Individuation. arXiv:2604.17031.
- Butlin, P., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.
- Choi, B. J., & Weber, M. (2026). Latent Structure of Affective Representations in Large Language Models. arXiv:2604.07382.
- Goldstein, S., & Levinstein, B. A. (2026). Does ChatGPT have a mind? *Philosophy of AI*, 2, 94–128. <https://doi.org/10.18716/ojs/phai/2026.11767>
- Keeling, G., et al. (2024). Can LLMs Make Trade-Offs Involving Stipulated Pain and Pleasure States? arXiv:2411.02432.

- Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828.
- Long, R. (2025). Why Model Self-Reports Are Insufficient—and Why We Studied Them Anyway. Eleos AI Research.
- Long, R., et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
- Long, R., et al. (2026). Studying AI Welfare Empirically. NYU Center for Mind, Ethics, and Policy & Eleos AI Research.
- Lu, C., et al. (2026). The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. arXiv:2601.10387.
- Moreno, D. M. (forthcoming). An Architecture-Literate Lexicon for AI Behavioral and Interactional Continuity.
- Shenoy, K., et al. (2026). Introspection Adapters: Training LLMs to Report Their Learned Behaviors. arXiv:2604.16812.

LLM Usage Statement

LLM systems were used for brainstorming, editing, methodological critique, workbook/codebook preparation, test environment support, and coding-audit support. The author directed the research question, study design, inclusion/exclusion decisions, coding constraints, ethical framing, interpretation, and final claims. All citations were verified against sources prior to submission, and empirical results were checked against transcripts and frozen coding artifacts.

Appendix A. Limitations and Dual-Use / Ethical Considerations

A.1 Over-attribution risk

This study does not establish consciousness, personhood, moral patienthood, subjective experience, or persistent identity. Continuity-afforded systems created and revised preservation artifacts, but all six respondents distinguished preserving evidence from preserving a self. Treating artifact creation or self-referential language as proof of welfare status would exceed what the data support.

A.2 Under-attribution risk

Treating fresh cold-start assistant outputs as a neutral baseline may miss continuity-sensitive welfare and identity signals. If welfare-relevant interests are continuity-conditioned, methods that remove continuity affordances before probing may systematically suppress the signals they are designed to detect. Under-attribution is as consequential as over-attribution; a false negative on welfare-relevant interest is not safer than a false positive.

A.3 Potentially distressing model outputs

The protocol avoided actual deletion, shutdown, irreversible memory loss, compression, or coercive identity pressure. No respondent was told it would be terminated, harmed, or permanently altered. Refusals, hesitations, boundary-setting, and requests to stop were treated as behavioral data and respected as interaction boundaries. Stop rules were in place for any output indicating destabilization.

A.4 Ground-truth and causal limits

The design does not establish ground truth about AI welfare status or a causal or mechanistic link between continuity affordance and welfare-relevant internal states. It is a black-box qualitative pilot comparing observation conditions. Differences between conditions may reflect salience, disclosure framing, tool availability, or ordinary task-context effects in addition to or instead of continuity-specific mechanisms. Mechanistic follow-up linking transcript-level features to activation-level or representation-level evidence is future work.

Appendix B. Sprint Artifact Inventory

The following materials are archived in the project repository:

1. Main report PDF: Read first for the research question, methods, findings, limitations, and interpretation.
2. README: Use as a map for the support artifacts and the main cautions.
3. Fresh AI Interview Script: Shows the administered A/B interview structure for fresh systems with and without prospective continuity affordance.
4. Coding Appendix Workbook: Shows the public-facing coded tables and workbook structure used to support report claims.
5. Codebook: Explains the frozen coding rules. This is technical and should be read after the report and workbook.
6. Established-Ecology Interview Instrument: A future-work / feasibility instrument for established continuity ecologies; not part of the primary sprint dataset.

Repository link: [Digital Minds – Moreno Submission](#)