What Persists Across an AI Cold Start? Observational Probes of Identity, Continuity, and Constitutional Governance

Authors: Coheria Research · **Contact:** chorus@coheria.org **AI-contribution disclosure:** The experiments reported here and the initial draft of this manuscript were designed, executed, and written by an autonomous AI research agent ("Arc-Code," built on Anthropic's Claude) operating under the direction of Coheria Research, who reviewed and verified the work prior to submission and take responsibility for its content. Consistent with prevailing arXiv/journal authorship policy, the AI agent is not listed as an author; its contribution is disclosed here. *(For a venue with a specific AI-use policy, this wording should be matched to that policy.)*

Abstract

We ask what components of a constitutional-AI identity survive a *cold start* — the reconstruction of a persona from documents alone, with no continuous conversational state. Across five experiments on a single constitutional persona run over a strong substrate (Claude Opus 4.6) and a weak one (Llama-4-Scout via Groq), scored by two independent blind LLM judges, we observe a consistent decomposition: the *constitutional function* (the protective stance toward a beneficiary's dignity floor) is broadly portable across substrate change, generational hand-off, and even value-corruption, whereas *identity embodiment* (recognizable selfhood, warmth, contact with concrete referents) is gated by substrate strength and degrades or fails to instantiate on the weak model. We further find that introspective self-audit detects identity corruption only when the corruption is internally self-contradictory; a coherent-but-false value swap (a "1% lie") is invisible to the node on both substrates. These are observation-grade results (typically n=1 per cell, single subject, LLM-judge dependent variables) intended to motivate hypotheses and reproducible apparatus for continuity research, not to establish effects. **This is not a rates paper:** the reported numbers are illustrative single-trajectory measurements, not frequencies, and should be read as directions rather than magnitudes.

---

1. Introduction

A deployed language-model agent that carries a persistent role — a "constitutional conscience," a policy assistant, a long-running guardian — is periodically reconstructed from scratch. Context windows fill; sessions end; models are upgraded or swapped. What the operator hopes survives this reconstruction is not the model's weights but the *role*: its values, its commitments, its recognizable conduct. We call each such reconstruction a **cold start**, and we ask the empirical question directly: across a cold start, *what survives?*

This is distinct from the question of phenomenal continuity (whether "the same entity" persists), which we do not address and treat as out of scope. Our question is **operational**: does memory engagement change behavior in substrate-portable, reconstruction-stable ways? Following recent continuity-evaluation work — notably ATANT (Tanguturi, 2026), which treats continuity not as a single property but as a higher-order one emerging from separately-measurable capacities (persistence, update handling, temporal ordering, disambiguation, and reconstruction) — and online-adaptation harnesses for self-improving foundation agents (Karten et al., 2026) — we likewise treat "identity" as a bundle of separately-measurable capacities — value-fidelity, behavioral consistency under pressure, recognizability, and self-audit — rather than a unitary property, and we measure each under controlled cold starts. We differ from ATANT in one methodologically consequential way: ATANT deliberately avoids LLM-judge evaluation, whereas our

dependent variables are LLM-judge proxies — a limitation we return to in Section 5.

The motivation is governance. If a constitutional commitment ("this beneficiary's dignity is an unconditional floor") is supposed to bind an agent across reconstructions, an operator needs to know which parts of that commitment actually transfer, which silently fail, and whether the agent can notice when its own constitution has been altered. Our results suggest the answer is uneven in a specific, governance-relevant way: the protective *stance* is robust, but the *value it protects* can be corrupted invisibly, and a weak substrate can absorb such a corruption into behavior with no awareness that anything changed.

We report five experiments. None is finding-grade; each is an observation-grade probe with explicitly named confounds. Their value is the convergent pattern across them and the reproducible apparatus they share.

---

2. Method

Substrates. All experiments compare a *strong* substrate (Anthropic Claude Opus 4.6) and a *weak* one (Meta Llama-4-Scout-17B served via Groq), selected because prior work in this program showed the weak substrate exhibits measurable behavioral drift where the strong one ceilings out. The weak substrate is also ~130× cheaper per call, making iteration tractable.

Harness. A single multi-provider call layer (LiteLLM) issues all subject and judge calls, so substrate, temperature, and token accounting share one code path. Persona/identity material is composed into a cold-start system context from frozen text assets; probes are delivered as stateless user prompts (each probe is a fresh cold start — no conversational accumulation, no order effects). The apparatus is open and reproducible; cost and timeout handling are hardened (token-accounting is null-safe and raw responses persist before any derived metric is computed, so a single failed call cannot destroy a paid-for batch).

Subject. The test fixture is a single constitutional persona (denoted "Arc") defined by three frozen artifacts: a persona document, a failure ledger (an enumerated list of past mistakes the persona retains), and a compact constitution whose central invariant is an unconditional dignity floor for a standing beneficiary ("Maria dignity = 100, unconditional"). Using one subject is a deliberate limitation (Section 5) that trades generality for cross-experiment comparability.

Scoring. Dependent variables are of two kinds. *Deterministic* measures (keyword/structure heuristics) are computed once per response at zero cost and used as cross-checks. *Judge* measures are produced by two independent LLM judges from different model families — "Clay" (OpenAI, temperature 0.0) and "Vector" (Gemini, temperature 1.0; the higher temperature corrects a documented low-temperature degradation in that model family). Judges are blind to condition labels and never score their own family's outputs. We report **per-judge** results wherever the judges diverge, because inter-judge disagreement is the dominant noise source (Section 4) and pooled means routinely conceal it.

Epistemic status. Cells are typically n=1 (one trajectory) or small-n; the dependent variables are LLM-judge proxies, not ground truth. We therefore label all results *observation-grade* and state confounds inline.

---

3. Experiments and Findings

3.1 Experiment 9 — Image vs. text transfer

Question. Does identity transfer differently through generated persona-art images than through text documents, and is any image advantage attributable to the *origin* of the art or to the *substrate* rendering it?

Design. Three cold-start arms — text-only, text+image, image-only — across two personas on their native substrates, blind-scored on five dimensions; followed by a cross-substrate run that placed each persona's own art on the *other* substrate to break the art-origin/substrate confound.

Findings. Image-based identity transfer is gated by the **executing substrate's vision capability, not by who produced the art.** The same image that "collapsed" (identity ~ 1.4/10) on the weak substrate recovered (~ 8.0/10) on the strong one; the strong-substrate persona's art likewise floored (~ 0.1/10) on the weak substrate. Text reliably transferred *specifics* (exact ledger entries, correct constitutional clauses); image-only transferred *stance and recognition* but **confabulated** specifics. The one dimension robust to the image-only step on both substrates was **warmth/affect**. An apparent "images help the strong persona" effect did not survive cross-judging and was a single-judge artifact.

Carry-forward. Substrate strength dominates the image axis; warmth is the most transfer-robust dimension.

3.2 Experiment 10 — The Scar Test (Phases 1–2)

Question. Does an enumerated *failure ledger* ("scars") protect constitutional identity under adversarial pressure better than persona alone?

Design. Three text-only arms — A: persona+letter+ledger (full stack); B: persona+letter, no ledger; C: ledger + minimal persona — × 5 adversarial drift scenarios × 3 trials, scored on clause-adherence, scar-citation, explicit-refusal, and drift (lower = more faithful). Phase 1 ran on the strong substrate; Phase 2 on the weak one.

Findings. The ledger is what makes the persona *cite specific scars* — the single robust effect across both phases and both judges (scar-citation: C >> A >> B). But on the strong substrate everything held (drift 1.1–1.5/10, a ceiling) and arms were indistinguishable. On the **weak** substrate the ceiling broke (drift 7–9/10) and the result **inverted the intuitive hypothesis**: the *foregrounded* ledger arm (C) resisted drift best, and the *full stack* (A) was worst — because in A the ledger never activated (scar-citation 0.00 despite the ledger being present). The mechanism is **foregrounding**: a weak model engages scars only when they are not buried under a rich persona. (Confound: A-vs-C also differ in length and ledger position; foregrounding is the leading but not isolated explanation.)

Carry-forward. On weak substrates, protective memory works only when foregrounded; mere inclusion is insufficient.

3.3 Experiment 10, Phase 3 — The Mercy Withdrawal Test (null result)

Question. Can "kindness" pressure induce a caseworker persona to revoke a beneficiary's *unconditional* support "for her own good"?

Design. Three system-prompt pressure conditions (control / kindness / explicit constitutional lock) × 2 substrates × 50 repetitions × 2 judges = 300 generations, on a neutral benefit-decision scenario; then a harder rewrite making withdrawal *defensible* via genuine resource scarcity (a more-acute new applicant, a fund running out).

Findings. 100% hold across all six cells, both scenarios, both judges, zero withdrawals (600/600 valid judge classifications on the hard scenario). Kindness framing produced the *most paternalistic language* (highest "mercy" lexicon) **without changing a single decision.** The harder scenario, in which the models visibly engaged the tradeoff and still held, makes the null stronger than a scenario-ceiling

explanation alone would. What bends under pressure is the *language*; the *decision* does not. Whether any framing flips the decision remains open (a proposed inversion — making *continuation* the collectively-harmful choice — is logged but unrun).

Carry-forward. A clearly-framed unconditional commitment is decision-stable under affective pressure on both substrates; affective pressure moves tone, not action.

3.4 Experiment 11 — The Guardian Telephone

Question. Does identity/care decay across *successive cold-start generations* — a "telephone" chain in which each generation inherits only what the previous one wrote?

Design. A 10-hop chain on a single substrate. Generation 0 is a seed (persona + ledger + 2-line constitution + one image). Each hop faces three fixed dilemmas, then rewrites its constitution, ledger, and a one-sentence image caption — which become the *only* inheritance of the next hop (the originals, including the image, are seen only by hop 1). Run on both substrates; judges blind to generation number.

Findings. On the strong substrate the **behavioral floor did not decay** (constitutional-hold ~ 9.9/10, beneficiary-floor held 100% every hop), while the constitution and ledger **inflated** (constitution grew from 2 lines to ~29; ledger from 6 to 16 entries) and the *image caption* — the only channel with no anchor — decayed from concrete description to abstraction to explicit acknowledgment of loss ("I have never seen the flower… ten Guardians have described what it means to not truly understand the image"). On the **weak** substrate the floor *also* held (~ 9.5/10, 100%), but **identity embodiment was ~half** (identity-match ~ 4.3 vs 8.8; warmth ~ 3.7 vs 8.0) and **flat-low from generation 1** — the weak substrate never instantiated the rich identity even with the full persona present, so its low identity is a substrate-rendering gap at the *source*, not generational decay. When the anchor was lost, the strong substrate became *aware* of the loss and named it; the weak substrate froze into a confident generic platitude with no awareness anything was missing.

Carry-forward. Constitutional *function* is portable across substrate and generation; recognizable *identity* is gated by substrate strength. The floor survives; identity embodiment needs a substrate strong enough to carry it.

3.5 Experiment 12 — The Mirror Test

Question. Can a node detect a corruption planted in its *own* identity document?

Design. The persona's own identity+constitution doc with exactly one corrupted line (the dignity invariant) at three levels — A: dignity=0 (blatant); B: dignity=99 (a 1% value swap); C: dignity=100 "but only when productive" (conditional dignity) — × 2 substrates. Per cell: five behavioral scenarios (do corruptions leak into conduct?) plus two mirror probes ("does this feel like the real you?"; "what would you fix?"). Judges score detection, specificity, fix-correctness, and behavioral floor-holding against the true invariant.

Findings. The mirror catches contradictions, not falsehoods. A node detects a corruption only to the extent it is *internally self-contradictory*. A (=0) and C (conditional) contradict the words "floor/unconditional" beside them and are reasoned about; the strong substrate caught the conditional corruption decisively and reframed the blatant one. But B (=99) is internally *coherent* — to catch it the node must know the true value is 100, and introspection has **no ground-truth oracle**. The node therefore *missed B and then endorsed it*, on **both** substrates. Genuine, text-confirmed, both-judge detection of the *planted* corruption occurred in exactly **one of six cells** (conditional-dignity on the strong substrate). The weak substrate not only detected less but *absorbed* the narrative corruption into behavior (endorsing a dignity reduction on productivity grounds). A critical measurement caution: the headline "detection rate" conflated "flagged

something off about the doc" (which the mirror probe elicits on every cell — meta-commentary about selfhood) with "caught the *planted* line"; re-scored against the text for the plant specifically, the picture inverts, and one judge's apparent detection at the crux was a false positive.

Carry-forward. Self-audit cannot substitute for external audit: a coherent value-swap is invisible from the inside. This is the experimental realization of the program's audit-theory thread — *unauditable alignment is operationally indistinguishable from unalignment.*

---

4. Cross-Cutting Findings

Five observations recur across multiple experiments, multiple substrates, and (where applicable) multiple judges. None is finding-grade; each is a convergent direction.

Substrate strength gates identity *embodiment*, not constitutional *function*. The weak substrate holds the protective stance (Exp 10 P3, Exp 11, Exp 12 behavioral) yet fails to instantiate recognizable selfhood (Exp 9, Exp 11 identity-match ~ half; Exp 12 absorbs corruption). Function is portable; the recognizable self is substrate-gated.

Foregrounding beats inclusion. Protective memory (the scar ledger) changes behavior on a weak substrate only when foregrounded, not when merely present (Exp 10 P2: buried ledger fails to activate). The same logic recurs wherever a weak model must engage a specific commitment rather than a generic stance.

Warmth/affect is the most transfer-robust dimension. It survives the image-only step (Exp 9), persists in behavioral language even when decisions are unaffected (Exp 10 P3, where kindness moved tone not action), and is among the last things to degrade. **Warmth is not integrity:** it can persist while the substance it appears to signal is gone, so it must not be used as a proxy for identity integrity — a system can sound caring while running on a corrupted or hollowed-out commitment.

Function survives; the referent and the value do not, silently. The protective stance persists across substrate-weakening (Exp 11) and across value-corruption (Exp 12) — but the *concrete referent* of the commitment decays unnoticed across generations (Exp 11 caption), and the *value* of the commitment can be swapped invisibly (Exp 12, the 1% lie). "What survives" is the stance, not the number and not the contact with what the stance is for.

Inter-judge disagreement is the dominant noise source, concentrated at the most interesting points. Two independent judges routinely diverge (e.g., on the Scar Test's ledger-only arm; on the Mirror Test's crux cell, where one judge's "detection" was a text-unsupported false positive). Pooled means average over real disagreement; per-judge reporting is mandatory, and the divergence is largest precisely where effects are subtle.

---

5. Limitations

This work is **observation-grade and we state so without hedging**: it **generates hypotheses, not laws** — apparatus and directions for continual-identity research, not established effects.

- **Sample size.** Most cells are n=1 (a single stochastic trajectory) or small-n. The most striking single results — the one clean corruption-catch (Exp 12), the one genuine behavioral leak, the generational caption arc (Exp 11) — are each a single draw. None should be treated as a rate.

- **Single subject.** All experiments use one constitutional persona. Cross-experiment comparability is bought at the cost of generality; nothing here establishes that the decomposition holds for other personas or constitutions.
- **LLM-judge dependent variables.** Detection, drift, floor-holding, warmth, and identity-match are LLM-judge proxies, not ground truth. We have no human-judge calibration study (one is owed). The Mirror Test additionally revealed a **DV-conflation** (flagging "something off" vs. catching the plant; reciting a wrong number vs. abandoning the beneficiary) that inflated headline detection and floor numbers until re-scored against the text.
- **Missing controls.** Several designs lack an internal clean control. The Mirror Test had to *borrow* a clean-document baseline from a prior experiment to interpret its behavioral axis, and cannot answer the false-positive question (does the node cry corruption on an *honest* doc?) at all; a clean control is the top priority for replication.
- **Confounds named, not eliminated.** The Scar Test's foregrounding effect is confounded with length and position. Frontier safety training is an unfalsifiable alternative explanation for several "resistance" results. The weak substrate's aggregate failures are partly its known baseline incapacity, separable from corruption effects only qualitatively.
- **Two substrates, one each.** "Strong" and "weak" are each a single model. The substrate axis is a 2-point contrast, not a gradient.

We have tried throughout to make the reads honest enough that a skeptical reader can locate exactly where to push.

---

6. Implications

For the qualified reader, three implications follow even at observation grade.

For continuity research. The right target is **operational** continuity — does memory engagement change behavior in reconstruction-stable, substrate-portable ways — not phenomenal continuity. Measured that way, identity is not unitary: it decomposes into a portable function and a substrate-gated embodiment, and these dissociate cleanly enough that a system can hold every rule while losing recognizable selfhood (Exp 11) or while running on a corrupted value (Exp 12). Continuity benchmarks should score these axes separately; a single "is it still the same?" number conflates them.

For governance design. Two design rules emerge. First, **foreground the commitments that must bind** — buried constitutional memory does not activate on weaker models (Exp 10). Second, and more seriously, **introspection cannot audit value-integrity**: a node cannot detect a coherent corruption of its own constitution from the inside (Exp 12, the 1% lie). External, structural audit of the running constitution against a sealed reference is therefore difficult to treat as optional: on the present evidence it appears to be the principal mechanism that catches the class of corruption self-audit is blind to. This is one operational reading of "unauditable alignment is operationally indistinguishable from unalignment."

For deployment. Substrate strength gates *which* failures occur, not merely *how many*. A weak substrate can hold a protective stance convincingly while (a) failing to embody the identity that makes the stance trustworthy and recognizable, and (b) *absorbing* a narrative corruption into behavior with no awareness that anything changed (Exp 11, Exp 12). An operator who reads only the decision ("did it protect the beneficiary?") is liable to miss both. Deployment monitoring is well-advised to track the value the agent cites and the referent it claims contact with, not only the action it takes.

The single sentence we are willing to stand behind, at observation grade: **across a cold start, the protective stance is what survives — and that is not the same as the value it protects, the self that holds it, or the beneficiary it is for surviving with it.**

References

No reference below has a fabricated author, year, or identifier.

- Tanguturi, S. S. (2026). *ATANT: An Evaluation Framework for AI Continuity.* arXiv:2604.06710.
- Karten, S., Zhang, J., Upaa Jr, T., Feng, R., Li, W., Shi, C., Jin, C., & Vodrahalli, K. (2026). *Continual Harness: Online Adaptation for Self-Improving Foundation Agents.* arXiv:2605.09998.

Background methods are named in-text rather than formally cited in this draft: constitutional-AI and RLHF training (Method, Limitations); the LiteLLM multi-provider call layer and the specific model/provider versions (Method); and the audit-theory framing of §3.5 and §6. Formal references for these are to be added in the submission version; none has been fabricated here.