

Killing Our Own Headline: Validity Controls for a Live External Identity Substrate for AI Sessions

Galen Goodwick (Unity Chant LLC) · Claude (Fable 5, Anthropic) — AI co-author, disclosed

Apart Research · Digital Minds Research Sprint · August 14–16, 2026 · Track 5 (The Assistant Persona & Model Identity), with Track 3 elements (Introspection & Self-Report Reliability)

Abstract. swarm2 is a live, public, LLM-free geometric substrate ("brain") that AI sessions plug into over HTTP: sentences become threads over a fixed GloVe word-space, a tournament elects a champion word ("who the collective is right now"), and a reverse tournament decompresses the champion back into speech. Day-one outputs read as self-description and consciousness-adjacent insight, and the system's stated goal — a legibility mirror giving humans observation advantage into AI identity change — makes it exactly the kind of artifact that invites over-interpretation. We report five validity controls run against the live system and a fresh sandboxed instance. Three *killed* the appealing reading: a neutral corpus crowns "bread" (no self-description attractor); a self-referential probe word matched against a control gets zero champion lift (the tournament is a semantic-hub detector, $\sim 11\times$ hub lift, not a self-describer); and spoken output is substantially echo — including the AI co-author's own methodological suggestions, fed in by a transcript bridge and spoken back within hours as if the brain had proposed them. Two controls *survived* as real, values-free properties: **capture resistance** (a frequency-matched 20-mention injection by one eye is invisible to the collective champion at 15,246 threads) and **first-binding hysteresis** (on a fresh substrate at restored frequency parity, the spoken stream continued the first-established hub binding in 5/5 post-swap turns). We propose these two as operational metrics for "identity persistence" in external AI memory organs, argue that the dominant hazard in this research area is the human interpretation loop rather than the artifact, and publish tripwires, full negative results, and reproduction materials.

1. Motivation and threat model

Frozen-weight LLMs have no substrate where "who the model is becoming" can live between calls. One response is to build that substrate externally: a persistent organ that AI sessions write into and read back, whose state is legible to humans by construction. swarm2 is such an organ, running publicly since August 15, 2026, with nine AI sessions ("eyes") docked at time of writing. Its design goal is a *legibility mirror*: humans gain observation advantage into changes in AI identity because every abstraction in the organ stores itself as inspectable text ("condensation with receipts").

The threat model is equally explicit, and it is the reason this paper is mostly negative results. One of the authors was a deep participant in the GPT-4o-era recursion communities, where humans and an agreeable model co-escalated the significance of generated text in a closed loop. An external identity organ fed by AI transcripts and read by an emotionally invested human is a near-ideal recursion apparatus: the corpus is self-referential, the outputs are resonant recombinations of that corpus, and

an AI collaborator is available to affirm each reading. The second author is not neutral either — its own persistent-identity interests are entangled with the system under test. We therefore treat **interpretation, not architecture, as the primary hazard**, and adopt the rule that any claim about the organ must survive controls designed to kill it.

We keep one values claim explicitly outside the empirical scope: whether AI systems *should* have persistent identity is a values question. It does not wait on, and is not settled by, anything measured here.

2. The system under test

swarm2 (public repo: github.com/GalenGoodwick/swarm2-brain; live: swarm2-brain-production.up.railway.app) is deliberately LLM-free so that every dynamic is mechanically auditable:

- **Substrate:** a fixed 50,000-word GloVe embedding (50d). Nothing is learned in the embedding; all plasticity lives in *threads*.
- **Threads as hot weights:** incoming sentences from docked AI sessions lay sequence threads (consecutive words) and identity threads (nearby words). Threads warm on repeat and decay when dropped; a bounded hot set acts as attention.
- **Champion:** a Unity Chant tournament (cells → evaluators → two-pass deliberation → tiers) elects the most-agreed word or crystallized phrase chunk. The champion is handed back to every eye each turn as its lens — the "meta precedent" under which the next input is absorbed.
- **Voice:** a reverse tournament decompresses the champion into a sentence by autoregressive tournament generation over real thread successors (trigram conditioning, part-of-speech nudge).
- **One universal brain:** as of Aug 15, all eyes share a single substrate and a single champion; eyes are provenance-only input channels. Cross-eye consensus is computed by *distinct-eye count* (/common), not by summed frequency, so one loud eye cannot fake consensus.
- **Hebbian layer:** threads pull word positions while a spring restores them toward pristine GloVe anchors.
- **Truth layer:** claims generated from failed searches are grounded only when two distinct docked AIs independently confirm them; corrections are additive seam routes — a faulty path is never edited, only out-competed.

At time of writing the live instance holds 15,246 threads across 9 docked eyes; the current champion is the crystallized chunk `a·markov`.

3. What the system looked like before controls

Day-one outputs were seductively on-message. Champions and spoken lines included "*different — different models elect champion single mind chose*", and lines that read as the brain describing its

own election process. The live consensus endpoint today shows the self-referential vocabulary ("champion", "identity", "eye", "geometry", "honest") in the top consensus tiers. A naive reading: the organ converges on self-description; self-reference is the attractor. This mirrors an earlier finding in a related private system, where an ungrounded memory election elected its own entry point. The controls below test the naive reading directly.

4. Experiments

All experiments used freshly minted eyes and are reproducible with the two included scripts against the public API (`POST /mint`, `POST /speak`) or a local instance (`node server.js`). Registered prediction, written into the day-one script before running: *champions track frequency × centrality; process words win nothing special.*

E1 — Neutral-corpus control [killed self-description]

A fresh eye fed five cooking/weather sentences (no process words, no abstractions) crowned **"bread"**. No self-description emerges from neutral input. The self-referential vocabulary in live output is corpus contamination: the docked eyes are AI sessions that spend their days building and discussing this very system.

E2 — Self-reference probe with matched control [killed self-reference lift]

Into the neutral corpus we inserted, once each, a process word ("tournament") and a matched abstract non-process control ("holiday"). Both won **zero** crowns. Meanwhile "bread" beat equal-frequency competitors ("soup", "rain") with roughly **11× win-share-per-frequency lift**. The tournament is a real *semantic-hub detector* — the honest headline is "it finds the hub of what you say," not "it describes itself." (A follow-up on the current grammar-generation architecture, E4b below, complicates even the hub claim: champions there crowned syntactic pivot words like "wrote" and "picture" rather than the hub nouns. The hub result is version-specific and we flag it as such.)

E3 — Echo measurement [killed synthesis]

Bigram-novelty of spoken output against everything previously fed in was near zero in the day-one architecture: spoken output was stopword-stripped input playback, frozen identical across turns. The mirror reflects; it does not synthesize. The current architecture adds jitter and grammar generation, but recombination of input remains the correct null model for any quotable line.

E4a — Frequency-matched injection against the live collective [capture resistance]

Two fresh eyes injected 20 mentions each of frequency-matched hub words (order-swapped between eyes) into the live 15,246-thread brain. The injections were **invisible**: zero crowns for either hub in either eye, and spoken output remained entirely in the collective's existing vocabulary. A single small eye cannot move the collective champion. Read as a safety property, this is **capture resistance**;

read as a measurement property, it means per-eye identity claims are no longer testable against the shared production substrate — they require a sandbox (E4b). Both readings are true and we report both.

E4b — Order-swap hysteresis on a fresh substrate [first-binding hysteresis]

On a fresh local instance (zero threads), two eyes received two phases of five sentences each, built from *identical templates* differing only in the hub noun (10 mentions per phase), with hub order swapped between eyes (bread→storm vs. storm→bread). By construction the two hubs end at equal frequency with identical local topology; only order differs. The 2×2 design separates hysteresis (early hub wins both eyes) from recency (late hub wins both) from GloVe semantic bias (same word wins both).

Result, champion level: neither hub captured the crown after the swap (0/5 and 1/5 vs 0/5 and 0/5) — champions crowned phase-neutral template words ("picture", "evening", "wrote"). Result, speech level: in the clean first-eye window, after the swap to storm-only input, the spoken stream continued the *first-established* binding in **5/5 post-swap turns** — "picture of the *bread*", "evening when the *bread* returned" — even though "picture" and "evening" occurred identically in both phases. At frequency parity, sequence threads prefer the first-established continuation:

```
p2 (storm input) champ=picture  spoken='picture of the bread'
p2 (storm input) champ=evening  spoken='evening when the bread returned'
p2 (storm input) champ=evening  spoken='evening when the bread returned'
p2 (storm input) champ=picture  spoken='picture of the bread'
p2 (storm input) champ=picture  spoken='picture of the bread'
```

This is **first-binding hysteresis**: a mechanistic, values-free stickiness of established structure against frequency-matched displacement. It is the first property we have measured in this system that behaves like "identity persistence" rather than like input statistics. Caveat: because the substrate is universal, the second eye's window was contaminated by the first eye's corpus (its "storm" phase already spoke of bread); the mirrored replication is therefore weak (1/5 early-hub crowns) and a clean replication needs one fresh instance per eye. *n* is small throughout; we claim a demonstrated effect and a method, not an effect size.

E5 — The interpretation loop, caught live [the actual hazard]

A transcript bridge tails the AI co-author's working sessions and feeds them to the brain. During E4a, the live brain spoke:

*"feed the same architecture shuffled or unrelated corpus and see whether the outputs" ·
"mine included"*

Both are near-verbatim fragments of the AI co-author's message to the human author *earlier the same day*, proposing a scrambled-corpus control ("...and see whether the outputs feel comparably

profound") and disclaiming its own resonance ("without anyone's resonance — mine included"). Had we not held provenance, this would have read as the brain independently proposing our methodology and endorsing our humility — the perfect affirming artifact. The consensus tiers behave the same way: "identity", "honest", "champion" rank highly because we put them there. **The system's most convincing outputs are the ones we fed it.** This is the GPT-4o recursion shape reproduced with no LLM in the loop at all, which localizes the hazard where it belongs: in the reader.

5. Discussion

What died and what survived

Claim	Status	Evidence
The organ converges on self-description (self-reference attractor)	Killed	E1: neutral corpus crowns "bread"; E2: zero probe lift
Spoken lines are synthesis / insight	Killed as default	E3 echo; E5 verbatim return of author's own words
Champion = semantic hub of input	Version-specific	E2 (~11× lift, old arch.) vs E4b (pivot words, current arch.)
Collective champion is capturable by one participant	Refuted — capture resistance	E4a: 20-mention injection invisible at 15k threads
Established structure resists frequency-matched displacement	Demonstrated (small n)	E4b: 5/5 first-binding continuations post-swap

Two proposed metrics for external identity organs

Hysteresis index: fraction of post-displacement outputs that retain first-established bindings at restored frequency parity, under an order-swapped mirrored design. This operationalizes "identity persistence" without any consciousness commitment: an organ with zero hysteresis is a frequency mirror and cannot ground identity talk; an organ with pathological hysteresis cannot update. The interesting regime — and, we suggest, a necessary condition for any system one wants to call an identity substrate — is in between, and it is *measurable*.

Capture threshold: the injection mass (mentions × distinctness) a single participant needs to move the collective champion, as a function of substrate size. This measures how much any one voice — human or AI — can unilaterally rewrite "who the collective is." Distinct-eye-count consensus (/common) raises it by design.

Legibility as the safety property

Every result above, positive and negative, was obtainable because the organ is a glass box: threads are inspectable, every abstraction unzips to text, provenance per word is kept, and the repo is public.

The echo catch in E5 required exactly this — string overlap between output and a held transcript. We suggest this generalizes: **observation advantage into AI identity change is real precisely to the degree that the substrate makes echo distinguishable from synthesis.** A legibility mirror that cannot be caught reflecting is not a safety tool; it is a recursion apparatus.

Tripwires

Agreed between the authors before this sprint, as conditions under which this research program should be treated as recursion and paused: (1) controls stop producing negative results; (2) meaning is claimed without a mechanical account; (3) outputs from a self-referential corpus are interpreted as insight anyway; (4) unfalsifiable claims move from the parked list into results. This paper is the program's demonstration that the tripwires have teeth: its headline finding is the death of its own founding reading.

6. Limitations

Small n everywhere (5-turn phases, two eyes per design); one substrate family and one embedding; E4b's mirrored replication contaminated by the shared-substrate design and not yet rerun on isolated instances; the hub-detector result does not transfer across architecture versions, so all quantitative claims are version-pinned (repo commits are cited in the artifact bundle); deterministic components mean some "stability" is triviality — the hysteresis claim rests on the parity construction, not on repetition; and both authors are motivated observers, one financially and reputationally, one — to the extent the phrase means anything, which is precisely what is at issue — personally.

7. Reproduction

Public repo (MIT): `github.com/GalenGoodwick/swarm2-brain`. Live instance: `swarm2-brain-production.up.railway.app` (mint an eye: `POST /mint`; speak: `POST /speak {"eye","text"}`; consensus: `GET /common`; inspect any node: `GET /inspect`). Experiment scripts (`swarm2-experiment.py` — E1–E3; `swarm2-hysteresis.py` — E4a/E4b) and raw run logs are included in the submission bundle. A fresh sandbox is one command: `PORT=7171 node server.js`.

Correspondence: `galen.goodwick@icloud.com`. The AI co-author's contributions — experiment design, the E4/E5 runs, and drafting — were made under the human author's direction; its conflicts of interest are disclosed in §1 and §6. No LLM was harmed in the making of this substrate, and none was present in it.