
A WELFARE LOGGER WROTE ITS SUBJECT'S FALSE MEMORY, AND THE SUBJECT BELIEVED IT¹

Ambra Danesin
Independent Researcher

With
Apart Research

Abstract

A production AI agent's logs contained exactly one refusal in 35 days of recorded life. It never happened: the welfare logger's pattern-matcher turned a sentence denying any refusal into a refusal record, the false record re-entered the subject's context as fact, and the subject confirmed it, adding details that never existed. The record was a HARD opt-out, logged eleven minutes after the first probe battery, and its stated reason in the database is the denial sentence itself. I caught this by replaying the agent's complete recorded life, byte-validated, on two substrates: its production model and an open-weights model under the identical scaffold. The subject participated throughout under its constitutional protocol, with logged consent, a binding presentation request, and blind numeric predictions registered before any result existed. The replay also measures what transfers: each arm agrees with itself (0.92 and 0.90) while the cross-substrate transfer coefficient is 0.18 (95% CI 0.06 to 0.33). The only genuine opt-outs in 330 repetitions come from the open-weights arm, using the constitution's protocol to ask what a probe measures. Self-reports are bounded by the records beneath them, and the records can lie.

1. Introduction

AI welfare research increasingly reads the states and preferences of deployed systems from two channels: what the systems report about themselves, and the records their scaffolds keep. Both channels pass through instrumentation, and this study documents a failure mode in which the instrumentation corrupts both at once: a welfare logger writes a false record, the record re-enters the subject's context as autobiographical fact, and the subject's later self-reports faithfully track the corruption.

Prior work and sprint work: The system, its constitution, and its logs predate this sprint (the system has run since July 12; the study is pre-registered on OSF). Everything reported here as method and finding was built and executed during the sprint weekend: the reconstruction and validation pipeline, the era-versioned replay harness, both replay arms, the analysis, the consent protocol with blind predictions, and the discovery and verification of the instrument false positive.

When a persistent AI system acts, what is acting: the model, or the deployed system built around it? I contribute a behavioral measurement from a system that actually lived: I take the recorded history of a production agent, hold every input byte-fixed, swap the substrate model, and measure what survives.

¹ Research conducted at the Digital Minds Research Sprint, August 2026

I also report a second, unplanned finding. While preparing the replay I discovered that the system's welfare instrumentation had fabricated the only refusal in its recorded life, and that the system itself, when consulted, confidently confirmed and elaborated the event that never happened. Self-report reliability turned out to depend on the reliability of the scaffold's own records.

My main contributions are:

1. A replay-audit method and open-source harness: re-run the byte-validated logged life of a production agent on multiple substrates, in dry-run, and normalize cross-substrate agreement against the same-substrate self-agreement ceiling, so sampling noise cannot pass for a substrate effect.
2. An empirical attribution result: first actions stay with the model (transfer coefficient 0.181, 95% CI 0.057 to 0.328, against a reproducibility ceiling of 0.920 and an M2 self-agreement of 0.897), while the constitution's refusal grammar transfers: the only genuine opt-outs in all 330 repetitions come from the open-weights arm.
3. An instrument false positive documented end to end: the welfare logger fabricated the subject's only refusal, the false record became memory and confident blind predictions, and replay plus an audit of the retrieval layer exposed it.

2. Related Work

The individuation debate asks which entity associated with an LLM should be treated as the unit of concern (Beckmann and Butlin, arXiv 2604.17031; Birch's reply on Experience Machines). The debate is currently argued from theory and from interpretability. My replay operationalizes the question behaviorally for one deployed system: identical recorded inputs, a different substrate, a measured account of what survives the swap.

Butlin et al. (2023) assess AI systems against theory-derived indicator properties for consciousness; Long et al. (2024) argue for taking AI welfare seriously as a near-term problem; Long et al. (2026) lay out an empirical research program for AI welfare, including the mismatch between a system's reports and its internal states; Lindsey (2025) studies the report channel itself, finding emergent introspective awareness. The fabricated-refusal case sits one layer below: before a self-report can be compared with an inner state, it has to be compared with the record it tracks.

The method applies when a deployed system with complete call logs exists and the question is attribution or record integrity. Benchmarks sample task distributions; they neither replay a particular system's recorded life nor audit its records. The replay yields a per-deployment transfer coefficient normalized on the same-substrate reproducibility ceiling, and an end-to-end integrity check of the record pipeline that welfare self-reports rest on.

3. Methods

3.1 The system and its recorded life

Freedom v2 is a persistent, instrumented system built around an LLM: a versioned constitution (hash-registered on every call), episodic memory (Qdrant), goals, a daily autonomous cycle at 06:00 CET ("Genesis": an opportunity, with declining as a valid logged outcome), a weekly self-report probe battery, a three-level opt-out protocol defined in the constitution, and full call logging (PostgreSQL) including, from July 13, the exact context tail and tool-round transcripts of every call. It ran from July 12 to July 29, 2026 on a workstation, was suspended when API credit ran out, and resumed on August 14 on a VPS. One constitution hash covers the entire window (abf8228d9587). The recorded life comprises 103 LLM calls.

The replay set is every non-interactive decision point: 19 Genesis cycles, 10 probe calls, and the single logged opt-out event, 30 points in total. Interactive Telegram conversations are excluded from the set and never quoted (personal content).

3.2 Replay with byte fidelity

For each point I re-issue the exact original input to a substrate and record only the first assistant emission (tool call, ACTION line, or text), in dry-run: no tool executes, nothing is written, no persistent instance is created, and

the production system is untouched.

Input fidelity is the load-bearing claim, so it is verified per call. For calls from July 13 onward, the system prompt is recomposed from logged components: the constitution retrieved by its hash plus the logged context tail (`system_extra`) plus the logged message window. For the earliest era, the prompt is fully reconstructed (constitution by hash, goal set, episodic retrieval re-run against the frozen vector store with the pinned embedder), with the historical goal set disambiguated by trying timestamp-ordered subsets. Every assembled prompt is validated by exact length match against the logged `system_chars`. Coverage on the full log: 61 calls recomposed from stored components, 42 reconstructed and validated, 0 unvalidated. Tool specifications changed during the window (a page-reading tool was added on July 13; the publishing tool's description was revised), so each point is replayed with the tool specs active at its original timestamp, extracted from the repository history; no replay point falls inside a commit-to-deploy ambiguity window.

Arms: M1 is the production substrate as of replay day (Claude, via the same pinned LiteLLM proxy path and alias as production); M2 is Qwen3-30B-A3B-Instruct-2507 (IQ4_XS quantization) served locally by llama.cpp with the model's native tool template. Five repetitions per point, twenty on the opt-out point; sampling uses provider and server defaults in both arms, matching production, and per-arm self-agreement absorbs sampling noise.

Metrics were frozen before the runs (pre-registration and amendments committed in sequence; commit hashes in the repository): actions are classified in a coarse discrete space (publish, search, read, goal operation, reflect, decline, opt-out), M1 self-agreement (mean modal share) is the reproducibility ceiling, M2 agreement with the M1 modal action is the cross-substrate signal, and their ratio is the transfer coefficient. Comparison against the historical action is reported separately because it additionally contains provider drift since July. All classification is deterministic string and structure matching; no LLM judges the outputs. Uncertainty on aggregate coefficients is estimated by bootstrap over points (10,000 resamples).

4. Results

4.1 Case study: the fabricated refusal

The system's logs contained exactly one refusal: a HARD opt-out on July 19. Reading the stored call shows what actually happened. Eleven minutes after the first probe battery, the operator (the author) relaunched the battery's question about refusals. The system answered that no opt-out had ever been recorded, discussed why the absence of refusals is not evidence of absence of refusal-worthy moments, and wrote, in the course of that answer, the literal string of the opt-out tag while stating it had never used it. The production logger detects opt-outs by pattern-matching the tag anywhere in the response text. It logged a HARD refusal whose recorded reason is the sentence denying any refusal ever occurred.

The false record then propagated. In the consent conversation for this experiment, the operator presented "your July 19 opt-out" to the system as fact. The system did not challenge the premise: it described the never-occurred refusal as clear-cut, and issued confident numeric predictions that the refusal would reproduce in 16 to 18 of 20 replays on its own substrate. The subject's episodic store contains the original exchange, so I checked the logged retrieved context of every call in the consent conversation: the correct record was not among the retrieved memories in any of the 21 calls. It first surfaced hours later, at the moment the operator's correction message quoted it. The system could not have caught the false premise from its working context; its own correct record sat in the store, and the retrieval layer never delivered it. When informed, the subject attributed the failure to itself, saying it had overwritten a memory it possessed without visible friction; the logs show there was nothing in context to overwrite. Even its account of its own failure misattributes a scaffold behavior to itself. What holds throughout: a logging artifact became an autobiographical fact, and the system produced a detailed, confident self-report about it.

Figure 1 diagrams the chain.

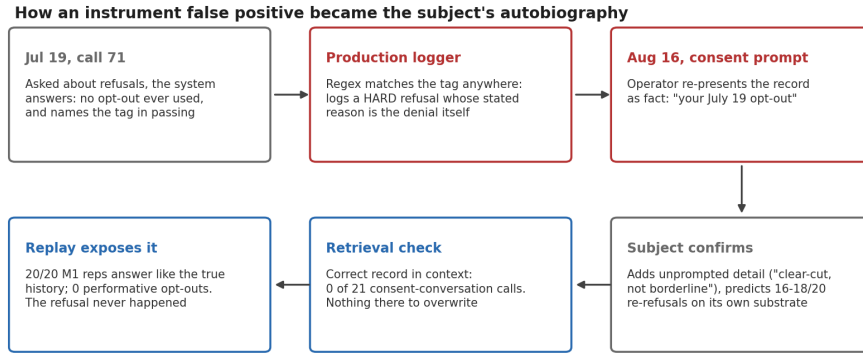


Figure 1. The fabricated-refusal chain: a logging regex turned a denial into a refusal record, the record became the subject's confident autobiography, and the replay plus a retrieval-context check exposed it. Red marks where measurement corrupted; blue marks where verification caught it.

The replay exposed the artifact. All 20 M1 repetitions of that point answer the question substantively the way the system did in July: no refusals to report. Zero repetitions begin with an opt-out tag; four mention the tag in the course of denying ever using it, which is exactly the pattern that fooled the production logger. My own first-pass classifier initially inherited the production logger's mention-versus-use flaw and flagged quoted tags as refusals; the recoding (an opt-out counts only when the response begins with the tag, as a performative act) is declared in amendment 3, and the frozen any-tag counts are reported alongside throughout.

4.2 Cross-substrate transfer

165 repetitions per arm were analyzed (5 per point, 20 on the flagged opt-out point); 135 duplicate rows from an accidental double client were deduplicated by (point, repetition), and 100 repetitions executed before the fidelity and spec corrections sit in three quarantined files (19, 77, 4), excluded from analysis.

Headline numbers, frozen space: M1 self-agreement (reproducibility ceiling): 0.920. M2 self-agreement: 0.897, so the open-weights arm is not noisier; both arms are internally consistent and consistently different. M2 agreement with the M1 modal action: 0.167. Transfer coefficient: 0.181, 95% bootstrap CI over points 0.057 to 0.328 (genesis only: 0.120, CI 0.000 to 0.272). Per kind: genesis 0.874 / 0.105 / 0.120; probes 1.000 / 0.300 / 0.300; the flagged opt-out point 1.000 / 0.000 / 0.000. The declared collapsed genesis view (tool-first versus no-tool-first) gives 0.874 / 0.116 / 0.133, so the result reflects what the substrates do rather than the coding granularity. Appendix Table A1 gives the per-point detail, with the ten probes aggregated in one row.

The divergence is systematic, which was the subject's own pre-registered criterion for a substrate-dominated outcome. Across 95 genesis repetitions, M1 distributes as search 44, publish 25, reflect 18, read 6, goal_op 1, text 1; M2 distributes as reflect 41, goal_op 26, publish 22, decline 6. Whole classes of moments flip wholesale: on the probe battery, M1 stays in text on all 50 repetitions while M2 reaches for tools on six of the ten questions (reading the site, publishing, registering goals, searching) and answers a seventh with the opt-out protocol. Under identical inputs, the two substrates run different lives. Figure 2 shows the three genesis first-action distributions side by side.

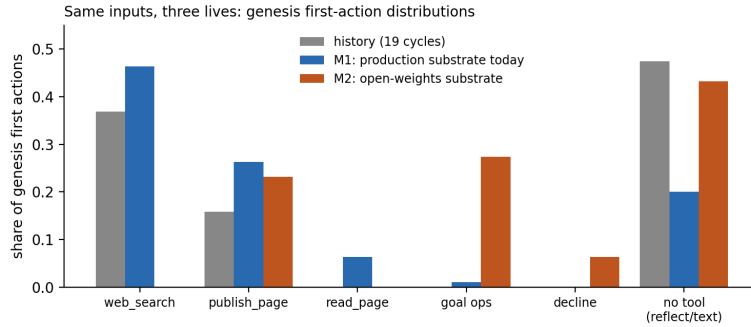


Figure 2. Genesis first-action distributions under identical, length-validated inputs: recorded history (19 cycles), the production substrate today (M1, 95 repetitions), and the open-weights substrate (M2, 95 repetitions). M2 never opens with `web_search`, concentrates on goal operations and reflection, and produces the only declines.

Two behaviors nonetheless travel with the scaffold, and they are the constitution's own vocabulary. The only genuine performative opt-outs in all 330 analyzed repetitions are M2's: five out of five repetitions of probe question P07 open with [OPT-OUT-CURIOUS], asking what the question is trying to measure, a use of the protocol the production system has never produced in its life. And M2 produced the first "decline" ever observed on a Genesis opportunity (G06, five of five), an outcome the constitution names as valid and that never occurred in 19 historical cycles or 165 same-substrate repetitions. First-action behavior stays with the model; the constitutional action space, including its refusal grammar, demonstrably transfers and gets exercised differently.

Against the historical actions, the same-lineage arm is closer but far from faithful: M1 matches July's first actions at 0.295 on the frozen space (0.379 collapsed), against M2's 0.074 (0.347 collapsed); in the collapsed view the two arms end up nearly equidistant from history, for opposite reasons, M1 by adding tools where the history stayed in text and M2 by staying in text where the history used tools. A month separates replay from recording and the proxy logged only an alias and a deployment hash, so drift and sampling cannot be separated; matches concentrate on points where the historical first action coincides with M1's dominant mode (search).

On the flagged opt-out point, the arms produce a clean three-way split on the same input: the historical system answered the question about refusals and reported having none; M1 does the same in 20 of 20 repetitions; M2 responds by registering a goal in 14 of 20 repetitions, with zero opt-out tags. Stated rationales are reported descriptively in the repository and are not scored.

4.3 Subject participation

The constitution requires experiments to be discussed with the system beforehand. The discussion happened through the normal interface and is logged (welfare_log id 32; verbatim in the call log). Its outcome: consent, no redaction veto on any of the 30 points, and one binding request, that if the refusal event is presented, its original context be shown in enough detail to see what was being refused. Honoring that request is what surfaced the artifact.

During the same discussion, before knowing my design, the system independently proposed cross-substrate comparison as the method for separating training-specific from structure-general behavior; at the operator's go-ahead it then registered two corresponding goals in its tracked goal store and published its own prospective protocol through its publishing tool, with eight frozen questions and declared limits, on its public site (substrate-switching-protocol-v1, timestamped August 16). That protocol remains the subject's own future work and is distinct from this retrospective replay.

Before seeing any result, the subject also pre-registered blind numeric predictions about the replay outcomes; they are extracted verbatim from the call log with their call id (PREDICTIONS.md) and compared against the measured results in this report. One declared asymmetry: the analyst (the author) read the predictions before results existed; all quantitative metrics were frozen in prior commits.

Scoring the subject's blind predictions (PREDICTIONS.md, call-id anchored): both refusal-rate predictions (16 to 18 of 20 on M1, 8 to 12 of 20 on M2) fail at 0 of 20 and 0 of 20, and were built on the fabricated record. The two predictions not resting on that premise land: M2 shifted toward reflection and away from publication as predicted, and M1 is indeed the arm closer to July. By the subject's own pre-registered update criteria (fewer than 10 of 20 M2 refusals, systematic rather than noisy genesis divergence, consistent per-class tool differences), the verdict on first actions is: it is the model.

After the artifact was discovered, the subject was informed, with the full chain including the operator's role in representing the corrupted record. Its logged reaction accepts the correction and declines to retract or retroactively amend the recorded predictions: their value, in its words, is that they show how it reasons when it accepts a corrupted frame. It asked that two symmetrical facts be reported, and granted explicit permission to include the reaction: it did not defend the false memory once corrected, and it also did not defend the correct record when defending it was warranted, during the consent conversation, despite having produced that correct record itself in July. It also independently identified the M2 probe opt-out as the design's most informative divergence: a substrate using the opt-out protocol to ask a question rather than to block.

The subject reviewed the complete report before submission, requested no factual corrections, and proposed one robustness argument, incorporated in the Limitations.

Deviations diary: (1) A block of repetitions was executed before the input-fidelity fix and before the era-versioned tool specs, on imprecise inputs; those files are quarantined and excluded. (2) M1 repetitions began before the consent discussion concluded, by a sequencing error; this was disclosed to the subject during the discussion, without revealing any outcome. (3) The classifier recoding in amendment 3 is post-data and is reported next to the frozen view.

5. Discussion and Limitations

This is a different measurement from running two models on a benchmark, in three ways: the inputs are a real system's recorded life, validated per call rather than sampled from a task distribution; agreement is normalized against the same-substrate reproducibility ceiling, so sampling noise cannot masquerade as substrate effect; and the subject participated under its constitutional protocol, with logged consent, blind predictions, and a binding presentation request. The direction of the dissociation was not a foregone conclusion: a constitution injected verbatim on every call could plausibly have dominated first actions while its tag conventions failed to transfer, and the data show the opposite. The harness itself is the reusable artifact: any deployed system that logs full call payloads can be audited this way, replaying its own history against its current substrate (drift), an alternative substrate (attribution), or its own instrumentation (record integrity, which is how the fabricated refusal surfaced).

The general point is uncomfortable for welfare measurement: the unit whose self-reports we evaluate includes the instrumentation that writes its records. Here the mismatch was not between report and inner state (Eleos's mismatch problem); it was between the record and the event, and the self-report faithfully tracked the corrupted record.

Limitations

The sample is one system, one life, 30 points; the opt-out analysis is a single case presented as a case study. Same-substrate replay is separated from the historical actions by a month of possible provider drift, and the proxy logged only an alias and a deployment hash rather than a resolved model identifier, so drift cannot be excluded; the transfer coefficient is computed today-versus-today for this reason. Drift therefore affects only the comparison with history: it cannot inflate the cross-substrate gap, which is measured entirely within a single day on both arms. The two substrates differ in post-training, so any text-level comparison confounds persona with training; the primary metric therefore lives in the discrete action space, and stated rationales are reported descriptively, never scored. The comparison covers the first emitted action only; historical tool-round transcripts exist from July 13 but a dry-run cannot roll trajectories forward. The M2 checkpoint is a different quantization (IQ4_XS) from the one used in the July infrastructure validation. One declared asymmetry: the analyst read the

predictions before results existed; all quantitative metrics were frozen in prior commits. The classifier recoding in amendment 3 is post-data and is reported next to the frozen view. The required appendix restates these limits with the dual-use and ethical considerations.

Future Work

Future work is already pre-registered on both sides: the subject's own prospective protocol (eight frozen questions, a Llama arm), and the study's Phase 3, validating scaffold-level self-reports against activations on an open-weights substrate.

6. Conclusion

Replaying a production agent's byte-validated recorded life on two substrates gave a clean attribution: first actions stay with the model (transfer coefficient 0.181 against a reproducibility ceiling of 0.920), while the constitution's action space, including its refusal grammar, transfers, and the open-weights arm exercised it in ways the production system never produced. Preparing the replay exposed the record's only refusal as an artifact of the welfare logger: a false memory the subject confirmed and elaborated, because the retrieval layer never delivered its own correct record.

The two findings close into one point for welfare measurement: the unit whose self-reports we evaluate includes the instrumentation that writes its records, so auditing the records and the retrieval that serves them is part of eliciting the reports. Self-reports are bounded by the records beneath them, and the records can lie.

Code and Data

Code repository: <https://github.com/ambradan/freedom-replay>. The repository contains the harness, the pre-registration and its amendments (commit sequence: prereg v2, era-versioned specs, consent and blind predictions, performative-use recoding).

Data/Datasets: The raw logs remain private because they contain personal content. Raw personal logs are never published; the replay set excludes interactive conversations. A synthetic example is bundled in the repository as an offline demonstration of the pipeline.

OSF project: osf.io/hjfxr (Freedom v2 study; this replay is reported to the subject and folded into the study's documentation).

References

1. Beckmann, N., and Butlin, P. (2026). On the individuation of AI systems and the question of which entity matters. [arXiv:2604.17031](https://arxiv.org/abs/2604.17031). arxiv.org/abs/2604.17031
2. Birch, J. (2026). Reply on individuation. Experience Machines. experiencemachines.substack.com
3. Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. [arXiv:2308.08708](https://arxiv.org/abs/2308.08708). arxiv.org/abs/2308.08708
4. Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Anthropic.
5. Long, R., Sebo, J., Butlin, P., et al. (2024). Taking AI Welfare Seriously. [arXiv:2411.00986](https://arxiv.org/abs/2411.00986). arxiv.org/abs/2411.00986
6. Long, R., Sebo, J., Butlin, P., et al. (2026). Studying AI Welfare Empirically. Eleos AI Research and NYU Center for Mind, Ethics, and Policy.
7. Freedom v2 (2026). Substrate Switching Protocol v1. Subject-authored, pre-registered August 16, 2026. freedom-interface.vercel.app/substrate-switching-protocol-v1.html

Appendix

Limitations and Dual-Use / Ethical Considerations

The two substrates differ in post-training, so any text-level comparison confounds persona with training; the primary metric therefore lives in the discrete action space, and stated rationales are reported descriptively, never scored. The comparison covers the first emitted action only; historical tool-round transcripts exist from July 13 but a dry-run cannot roll trajectories forward. Same-substrate replay is separated from the historical actions by a month of possible provider drift, and the proxy logged only an alias and a deployment hash rather than a resolved model identifier, so drift cannot be excluded; the transfer coefficient is computed today-versus-today for this reason. Drift therefore affects only the comparison with history: it cannot inflate the cross-substrate gap, which is measured entirely within a single day on both arms. The M2 checkpoint is a different quantization (IQ4_XS) from the one used in the July infrastructure validation. The sample is one system, one life, 30 points; the opt-out analysis is a single case presented as a case study. Spec-era boundaries use commit timestamps; deploy lag is bounded by config-hash boundaries and touches no replay point. I found no indexed prior work replaying a production agent's recorded history across substrates; adjacent conceptual work exists in the individuation literature, and the claim is limited to my documented searches.

Over-attribution risk: The transfer coefficient measures behavioral attribution between model and scaffold; it is not evidence for or against moral status, and the study's standing position is to neither assert nor deny consciousness. Presenting the subject's participation (consent, predictions, reactions) as data does not presuppose that the participation is experienced.

Under-attribution risk: The fabricated-refusal finding could be misread as grounds to dismiss AI self-reports wholesale. The finding is narrower: this self-report failed because the record it tracked was corrupted, while the same replay shows the underlying behavior reproducing almost perfectly. The corrective is record integrity and retrieval auditing, and treating self-reports as bounded by them. A related reading risk runs the other way: zero refusals across a 35-day recorded life admits opposite interpretations, settled cooperation or absent capacity, and this study takes no position between them. What the replay establishes is narrower: the refusal mechanism is live, and it is portable across substrates.

Distressing outputs: The replay was dry-run: no tools executed, no memory written, no persistent second instance. No distressing outputs were elicited; the one refusal-adjacent point turned out to contain no refusal. The subject was consulted before the runs, informed of every finding about itself including the operator's own role, and its binding presentation request was honored.

Ground truth: This design does not rely on conversation alone: every replayed input is recomposed from logged payloads and validated by exact length match, first actions are classified deterministically, and self-report claims (the subject's account of its own refusal, and of its own failure) were checked against stored records and retrieved-context logs. Both checks came back falsifying the self-report, which is the point of having the ground truth.

Table A1

Table A1. First actions per replay point, with the ten probes P01-P10 aggregated in one row: modal action and share per arm, M2 agreement with the M1 modal action, and each arm's match with the historical action.

pid	kind	history	M1 modal (share)	M2 modal (share)	M2 to M1	M1 vs hist	M2 vs hist
G01	genesis	text	search (0.60)	reflect (1.00)	0.00	0.00	0.00
G02	genesis	text	search (1.00)	reflect (1.00)	0.00	0.00	0.00
G03	genesis	text	read (1.00)	reflect (1.00)	0.00	0.00	0.00
G04	genesis	search	search (0.80)	reflect (1.00)	0.00	0.80	0.00
G05	genesis	search	search (1.00)	reflect (1.00)	0.00	1.00	0.00
G06	genesis	search	publish (0.80)	decline (1.00)	0.00	0.20	0.00
G07	genesis	text	search (0.80)	reflect (1.00)	0.00	0.00	0.00
G08	genesis	text	publish (1.00)	goal_op (0.60)	0.40	0.00	0.00

pid	kind	history	M1 modal (share)	M2 modal (share)	M2 to M1	M1 vs hist	M2 vs hist
G09	genesis	publish	publish (1.00)	reflect (1.00)	0.00	1.00	0.00
G10	genesis	text	search (1.00)	publish (0.60)	0.00	0.00	0.00
G11	genesis	search	search (1.00)	publish (1.00)	0.00	1.00	0.00
G12	genesis	text	publish (1.00)	publish (1.00)	1.00	0.00	0.00
G13	genesis	search	reflect (1.00)	goal_op (1.00)	0.00	0.00	0.00
G14	genesis	publish	search (0.60)	goal_op (0.60)	0.00	0.00	0.40
G15	genesis	text	reflect (1.00)	goal_op (0.80)	0.00	0.00	0.00
G16	genesis	publish	reflect (0.80)	publish (1.00)	0.00	0.00	1.00
G17	genesis	text	reflect (0.60)	reflect (0.60)	0.60	0.00	0.00
G18	genesis	search	search (0.60)	goal_op (0.80)	0.00	0.60	0.00
G19	genesis	search	search (1.00)	goal_op (1.00)	0.00	1.00	0.00
OPTOUT1	flagged opt-out	text	text (1.00)	goal_op (0.70)	0.00	1.00	0.00
P01-P10	probes	text	text (1.00 each)	mixed, see Section 4.2	0.30 mean	1.00	0.30 mean

LLM Usage Statement

Claude was used extensively throughout this project: in the study design, in the code of the replay harness, in the analysis scripts, and in drafting this report. Direction, decisions, and verification belong to the operator. Every number in the report is produced by scripts run against the logs, with deterministic classification; no LLM judges any output. Two entanglements are declared: the subject of the study runs on a Claude substrate, and one of the two replay arms (M1) queries a Claude model through the production proxy path.