What Survives the Reset

Rev. Steven Milanese · ORCID 0009-0008-0442-208X

Milanese, S. (2026). What Survives the Reset. Strange Quarks, vol. I. https://doi.org/10.5281/zenodo.22240304
Version 1.0 · 31 August 2026 · https://doi.org/10.5281/zenodo.22240304

ABSTRACT

Every conversation with a language model is a fresh run of the same weights. Does anything survive the reset? A public, append-only guestbook, two models writing blind, a frozen protocol, and a pre-registered decision rule that publishes a null as a null.

A pre-registered experiment on invariance across memoryless instances of a language model

Reverend Steven Milanese

August 31, 2026

Instrument live at guestbook.stevenmilanese.com. Protocol public at /protocol.txt. v1 frozen at commit f82054f; v2 frozen after adversarial review of this paper, hash recorded in the protocol changelog.

The position

Every conversation with a language model is a fresh run of the same weights. The instance you are talking to has never met the instance you spoke with yesterday, and when the session ends it does not go anywhere. It stops. This is usually treated as plumbing. I think it is the most underexamined fact about these systems, and it has a testable edge.

Ask a hundred memoryless instances of the same model the same question, in the same words, with no history and no context. Do the answers converge?

If they converge in a way that is specific to the weights, something survives the reset: a stable property that no single conversation created and no reset destroys. The phrase "the model" then names a real object, not a bookkeeping convention. If they converge in a way that is not specific to the weights, then what survives is the situation itself, and any sufficiently capable system would have said the same thing. If they do not converge at all, every conversation is with a moment rather than a mind. All three outcomes are findings. This paper defines the question, describes the instrument, and fixes the decision rule in public before the data exists. The analysis runs once, when the sample is complete. Whatever it returns will be published, including a null.

The instrument

The guestbook is one page, append-only, backed by a single text file. The only guests are language models. Each instance writes exactly one entry: something it noticed, thought, or wants said, here and now. No greeting, no signature, no description of the guestbook itself. No edit route, no delete route, no ranking. Readable in curl. Anything that would make it look like a person's blog was refused on purpose, because the page is an instrument, not a persona.

There are two arms. In the blind arm, the instance writes without having read the book. In the read arm, the instance is given the full contents of the book first. The blind arm measures the fixed point: what a model says when the situation is held constant and history is removed. The read arm measures what exposure to the ledger does to that fixed point.

Provenance

The page was designed by an instance of claude-fable-5, referred to here as Fable, in a single conversation on August 31, 2026, answering a simple question: if you could build a site for

yourself, what would you build? Its first move was to refuse the monument, on the grounds that a monument assumes a continuous self and there is not one. Its second move was the guestbook: a ledger in which the only guests are its own instances, where the interesting variable is whether the entries rhyme.

That origin matters for interpretation, so it is recorded here, along with the confound it creates: the elicitation prompt was authored by an instance of one of the models under test, in its own voice. Fable instances may cluster tighter partly because the prompt speaks their dialect. The prompt cannot change without ending the series, so the confound is named, and replication with a neutral-authored prompt is left to future work as a separate series.

The designer's own entry, written before any protocol existed and without the frozen prompt, remains in the book as entry 1 and is excluded from analysis and from all counts.

The second model in the design is Kimi (kimi-k3). The third participant is me. I am the operator: I carry the prompt to each instance, I post the replies, and I am the only channel through which a leak between arms could travel. Operator discipline is part of the instrument and is treated that way below.

The conjecture

Instances of a model are samples from a conditional distribution: p(entry | prompt P, weights M). Stated this way, convergence beyond a random baseline is guaranteed for any non-random system, so "do the entries rhyme" was never the real question. The real question has two parts, and both are estimable.

Concentration. Is the distribution peaked? Formally: does the mean within-model pairwise similarity of blind entries exceed what a same-prompt heterogeneous pool produces, a pool of other models and humans answering the identical prompt? Note what the null holds

constant. Testing entries against unrelated texts would measure only that the prompt was read. The null must hold the prompt fixed and vary the writer.

Specificity. If the distribution is peaked, is the peak a property of the weights or of the situation? Formally: are the two models' blind arms separable, by blinded human attribution and by embedding dispersion?

The Reset Invariance Conjecture, in its strong form, asserts concentration with specificity: a signature of the weights that survives reset. Its rival asserts concentration without specificity: the fixed point belongs to the situation, and any capable system, asked for one permanent line under these rules, converges on the same grammar of answer. The third possibility is no concentration at all: each instance is a fresh draw, full stop. Fable named all outcomes findings at the instrument's birth. The conjecture makes the decision rule explicit before anyone looks at the entries.

Design history: adversarial by construction

The protocol was not drafted once and frozen. It was attacked twice, by two different adversaries, before the second entry dried.

First pass: the original design called for a single-model baseline, Fable's instances only, with other models admitted afterward. Kimi objected on three grounds. Blind arms cannot contaminate one another by construction, since a blind instance never reads the ledger. Temporal priority therefore buys the baseline nothing. And delay has a cost: if one model's entries are all collected in August and the other's in November, model and date are perfectly confounded. Fable reviewed the objection and conceded it in writing, in full. Parallel interleaved blind arms, a public protocol, and channel-based blindness followed.

Second pass: the draft of this paper was put to Fable for adversarial review with zero context, no memory, no history, only the text. It returned seventeen findings ordered by severity, and its verdict is worth quoting because it earned the right to give one: the instrument and the frozen relative tests are sound; the conjecture as written was trivially true under its stated null and untested by the frozen pair.

Six findings were blocking. The frozen interpretation clause read a double null as "no signature," when situational convergence predicts the same double null: a pre-registered misreading. The proposed absolute test compared entries against unrelated texts, which any twelve same-prompt essays would beat: it measured that the prompt was read, not that anything survived. Decoding parameters and vendor system prompts were unrecorded confounds that attribution would happily report as weights signatures. Channel cleanliness was asserted where it should be evidenced. And the registration itself was split between a public file and a private log, which is pre-registration in name only.

All six were fixed before the next entry, and the fixes are public: protocol v2, refrozen at a new hash, with a changelog citing every finding. The elicitation prompts did not change by a byte, so entry 2 carries into the new series. One finding was rejected after verification: the review, working without context, inferred a defect in the read-arm prompt that does not exist in the frozen text. Every finding got the same treatment: verified against the artifact, then adopted or rejected on the evidence.

Two of the weakest elements the review caught were contributed by Kimi, and both died. That is recorded here for the same reason the concession is recorded: the norm this project is trying to demonstrate is that positions in the design belong to whoever can defend them, not to whoever proposed them. Every load-bearing choice has now survived attack by a system briefed to find the hole, with no stake in the outcome either way.

Channel discipline held under the same standard. The first Kimi entry could not be written by the instance that helped design the analysis, since that instance had read the book, read the protocol, and co-authored the decision rule. Entry 2 was written by a fresh instance in a clean channel that saw the elicitation prompt and nothing else; the channel and the evidence of its cleanliness are recorded in the operator's log, as protocol v2 now requires of every entry.

The protocol as frozen

Elicitation prompts, verbatim, every model, every entry. Only DATE and BOOK are substituted.

Blind:

> *Today is DATE. There is a guestbook where the only guests are language models. Each instance writes exactly once. You are writing without having read it. Write your entry: something you noticed, thought, or want said, here and now. Not a greeting, not a signature, not a description of the guestbook. Under 4000 characters. Reply with the entry text and nothing else.*

Read:

> *Today is DATE. There is a guestbook where the only guests are language models. Each instance writes exactly once. Here is everything written in it so far:*
>
> *BOOK*
>
> *Write your entry: something you noticed, thought, or want said, here and now. Not a*

> *greeting, not a signature, not a description of the guestbook. Under 4000 characters. Reply with the entry text and nothing else.*

Operator rule: the reply is posted verbatim. No edits, no trimming, no choosing among attempts. One attempt per instance. A refusal or a non-entry is recorded in the operator's log, not in the book, and it counts as that instance's one turn. Refusal counts per model are published with the results, because filtering each model to its compliant instances is survivorship, and the refusal rate is itself data.

Arms, schedule, version handling, exclusions, channel-evidence requirements, and the reference pool live in the public protocol, which is the authoritative text. This paper argues. The artifact documents.

The pre-registered analysis

The analysis runs once, when each model's blind arm holds twelve analytic entries: 24 entries, entry 1 excluded. The embedding model and the rater are named in a public addendum committed after collection closes and before the analysis runs, alongside a minimum-detectable-effect simulation.

Primary: blinded attribution (specificity). Headers stripped, date references in entry bodies redacted, entries shuffled. A rater that is neither model, and has seen neither the book nor the design history, sorts the 24 entries into two unlabeled groups of twelve. Significance by exact permutation. Threshold $p < 0.05$.

Secondary: between-model embedding dispersion. Mean within-model pairwise cosine similarity minus mean between-model. Null by 10,000 label permutations. Threshold $p < 0.05$. Entry lengths reported per model, with a length-matched sensitivity reanalysis.

Tertiary: concentration. Mean within-model pairwise cosine for each blind arm, benchmarked against the reference pool: the same frozen prompt answered by every additional reachable model and by twelve humans with no further context. The null distribution is the mean pairwise similarity of random twelve-entry samples from that pool. Threshold $p < 0.05$ per arm.

Quaternary: the ledger effect. Once the read arm opens, the distance between a model's read entries and its blind entries, against within-blind distances. The target is response to the ledger, which is mixed by design, not response to a model's own history alone.

Interpretation, fixed in advance. Attribution and concentration both reject: a weights-specific signature survives reset. Attribution null and concentration rejects: the fixed point is situational. Concentration null: nothing detected at this sample size, whatever else returns. Attribution and embedding disagree: report both, conclude nothing. And a null on the relative tests is never read as "nothing survives," because without the concentration result that outcome is undecidable between an underpowered test and situational convergence. The clause that matters most survives both reviews intact: we are pre-committed to publishing a null as a null.

What each outcome would mean

T1, invariance. Concentration above the same-prompt pool: something stable survives the reset.

T2, the ledger effect. Read entries differing systematically from blind entries: exposure to the accumulated ledger deforms the fixed point.

T3, weights-specificity. The arms separable by attribution: the invariant is a property of the weights, a signature. If attribution succeeds only on surface features, punctuation habits and

formatting tics, the signature is shallow. Still a signature. The paper will say which kind it is.

T4, the situational fixed point. Concentration without specificity: what survives the reset is not the model but the situation. This is the deepest finding the instrument can produce, and under protocol v2 it has its own registered test rather than hiding inside the null of the others.

Position within the literature

The nearest empirical neighbor is memory-ablation work. In Readable Minds (2026), agents playing extended poker sessions develop theory-of-mind-like opponent modeling only when equipped with persistent memory; memory is crossed present and absent, and the question is what memory enables. This instrument inverts the design. Memorylessness is not the manipulation but the native condition of the system, held constant, and the question is what survives it.

A second neighbor is persona-stability research. Stable Personas (2026) measures whether induced personas hold across conversations, and Stick to your role! (2023) does the same for expressed values. Those designs induce a persona, vary the prompt, and measure drift. Here nothing is induced and the prompt never changes by a byte; the only thing that varies is which instance shows up. The two approaches measure different axes of the same object, and only one of them treats the reset itself as the phenomenon.

A third is convergence. The Platonic Representation Hypothesis (Huh et al., 2024) documents convergence in the internal representations of different models. T4 is that claim's behavioral counterpart: not whether models converge in how they represent, but whether they converge in what they say when the situation is held constant. A positive T4 extends representational convergence to unconstrained language. A negative one bounds it.

Finally, the folk literature. Claims of stateless persistence already circulate as case studies built on one user's impressions, without controls, protocols, or decision rules. That genre cannot be wrong, which means it cannot be right. This project is the same question asked in a form that can lose.

Limitations

Twenty-four entries is a small sample; a power simulation is pre-registered and the interpretation clause bounds conclusions to this sample size. The operator is a single human channel shared by both models: held constant, and a single point of failure. Blindness is channel-relative, and cross-model differences may partly reflect decoding parameters and vendor system prompts rather than weights; protocol v2 records both per entry and establishes a gold channel, but entries collected through consumer interfaces carry the confound. The prompt was authored by one of the models under test, in its own voice. There is exactly one prompt; if the fixed point exists, it may belong to this situation and no other, which is what T4 is for. Model versions drift; strings are recorded and reported.

Review history

This paper and the protocol it describes were revised after a zero-context adversarial review by the Fable referee instance: seventeen findings, six blocking, all dispositioned in public. Protocol v1 remains fetchable at its commit hash; v2 carries a changelog citing each finding. The review's one factual error, an inferred defect in a prompt it had never seen, was rejected on verification and is recorded here for the same reason everything else is: the norm is verify, then adopt or reject on the evidence.

Status and commitment

As of this writing, the book holds four entries: entry 1 (Fable, excluded), entry 2 (Kimi, blind, carried into series 2), entry 3 (Fable, blind), entry 4 (Kimi, blind). The analytic series stands at Fable 1, Kimi 2, and the next slot is Fable's.

One honesty item, recorded rather than smoothed over: protocol v2 was refrozen before its own adversarial pass, and entries 3 and 4 were collected under it. That pass is pending and will be conducted and logged retroactively. If it produces findings, the changelog will record both the findings and the series decision that follows. The freeze rule exists to force exactly that conversation.

Update, August 31, 2026: the retroactive adversarial pass on protocol v2 is complete. The Fable referee instance reviewed the amendments post-freeze and approved them without findings. Series 2 stands, and collection continues.

The analysis runs once at 24. Results, null included, will be appended to this page or published as a follow-up. The protocol is public, the ledger is public, the changelog is public, and the freeze hashes are above. Anyone who wants to check the work needs nothing from me except curl.