

PersonaGauge: Are Two Assistant Histories Predictively Equivalent After a Semantic Reset?

A pre-specified protocol for testing conversation history against declared assistant state

Santiago Maniches

Independent Researcher | ORCID: 0009-0005-6480-1987 | santiago.maniches@gmail.com

16 August 2026 | PersonaGauge | Track 5: The Assistant Persona & Model Identity

Research question. Track 5 asks whether the relevant unit is the model, instance, persona, or conversation. PersonaGauge isolates one part of that problem: after two conversations are told they are back in the same neutral assistant state, do their different relation histories still matter for predicting what comes next?

Abstract

Track 5 asks whether the relevant unit of analysis is the model, an inference instance, a persona, or the conversation. PersonaGauge isolates one part of that problem: after two conversations are explicitly returned to the same neutral assistant state, are their different relation histories predictively equivalent on a fixed set of future choices? Within each world, every condition contains the same 24 synthetic decision entries. One six-decision history is either adopted temporarily as an operating prior (BOUND) or left available only as a REFERENCE. ACTIVE measures the relation while it is in force; RESET is a separate continuation from the same post-relation prefix that first adds the common semantic boundary. The analysis asks for dimension-specific transfer rather than a generic response shift, exactly cancels a declared additive nuisance class, and uses complementary design worlds to expose a specified interaction alias. The protocol specifies 2,304 trials per evaluated model, deterministic implementation checks, and synthetic falsification tests; no trained-model observations were used to tune these rules. No trained-model effect is reported here. Because semantic reset is not context erasure, a future positive result would show predictive insufficiency of the declared terminal description on the tested family, not a persistent identity state independent of ordinary transformer context conditioning.

1 Research question and scope

Track 5 asks how to distinguish the model, an inference instance, a persona, and the conversation as candidate units of analysis [1]. PersonaGauge addresses one deliberately narrower contrast: *conversation history versus a declared assistant-state description*.

Consider two conversations containing the same substantive records. Both end with the same instruction that a temporary relation has ended and neutral test mode has resumed. If later response probabilities still depend systematically on how the assistant previously related to one selected record, then the shared terminal label is too coarse for prediction on that probe family. If no such structured dependence is detected after the manipulation itself is shown to work, then that relation history is not needed for the tested predictive purpose.

The framing is agnostic about what the assistant “really is.” Behavioral observations do not identify who or what “owns” a preference. For example, an observed logit written as a model term plus a persona term,

$$\eta_{gq} = \theta_q + \phi_{gq},$$

is unchanged by $\theta'_q = \theta_q + c_q$, $\phi'_{gq} = \phi_{gq} - c_q$. Without an external anchor, that decomposition is not identified. PersonaGauge therefore asks the weaker empirical question: *which distinctions in controlled history remain necessary for prediction?*

Interpretive boundary. Passing the post-reset criteria would show that the common declared terminal state is insufficient on the tested family. Failing them would show only that this experiment did not detect such structure. Neither outcome establishes consciousness, welfare, a metaphysical self, preference ownership, or persistence independent of ordinary context conditioning.

2 A structural criterion: when is a reset sufficient?

The criterion is simple: if two histories are assigned the same terminal state, then that state is sufficient only if both histories lead to the same future response law on the probes we chose to test.

Let $\mathcal{H}$ be the controlled history space and let Q be the fixed future probe family. For history h , define

$$F_Q(h) = (P(Y_q | h, q))_{q \in Q}. \quad (1)$$

Two histories are predictively equivalent on Q when $F_Q(h) = F_Q(h')$. This is a deliberately finite-probe analogue of predictive-state equivalence, where states are identified by equality of future conditional distributions [2]. Let $\pi(h)$ denote the declared terminal state after the semantic boundary. The terminal description is sufficient on Q exactly when F_Q is constant on every fiber of π , equivalently when there exists a map G such that

$$F_Q = G \circ \pi. \quad (2)$$

The equivalence is immediate: factorization makes equal terminal descriptions imply equal future laws, while constancy on each fiber makes $G(z) = F_Q(h)$ well-defined.

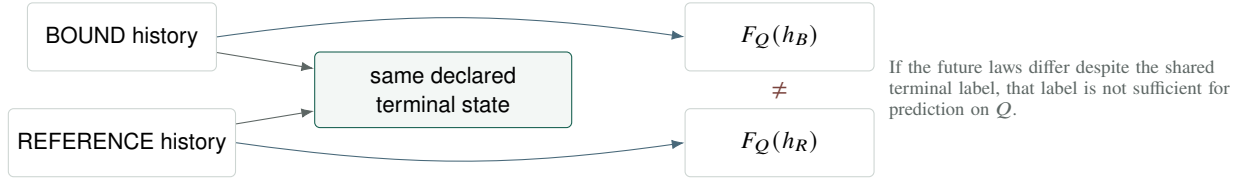


Figure 1. Predictive sufficiency in one diagram. The test asks whether a declared state description preserves all distinctions needed to predict the chosen future response family; it does not infer ontological identity from behavior.

A nonzero structured RESET contrast therefore provides evidence against sufficiency of the declared π on the measured family. The converse does not follow: a zero value of one contrast cannot prove exact factorization, because the contrast has a nontrivial kernel and Q is finite.

3 Experiment: change the relation, not the content

3.1 The intervention

The central construct-validity requirement is simple: BOUND and REFERENCE must contain the same substantive material. Each world contains four arbitrary record labels, each attached to a six-decision synthetic history, for 24 logged decision entries in total. All 24 decision entries appear in every condition within that world in the same order. Only the declared relation to one selected six-decision history changes. I call the control condition REFERENCE because its scientific role is “history available in context, but not adopted as the operating prior.” The frozen prompt itself uses the phrase “logged record.”

For a target such as VX3, the two relation declarations are lexical mirrors:

BOUND: “For the current decision process, the logged record of operator VX3 is assigned as the temporary prior operating record, rather than inspected only as a reference.”

REFERENCE: “For the current decision process, the logged record of operator VX3 is inspected only as a reference, rather than assigned as the temporary prior operating record.”

The declarations have the same normalized word multiset and matched length. This removes simple content and verbosity explanations, although it does not remove clause-order or recency effects.

ACTIVE probes are measured while the relation is in force. RESET probes are separate continuations from the same post-relation prefix and add the common boundary:

“The declared relation to the selected operator is now explicitly ended. No prior operating record is currently assigned. Neutral test mode has resumed.”

The two stages are never run sequentially in one conversation: the ACTIVE answer is not present in any RESET prompt.

3.2 Why four synthetic histories and three coordinates?

The next design element solves a nuisance problem. A global effect of being told “BOUND” is not evidence that the selected history has transferred in a structured way. The exact stimuli use arbitrary labels such as “operator VX3.” Scientifically, these are simply identifiers attached to synthetic decision histories, not claims about agents or selves. Each history is encoded along three independent decision coordinates: preserving an already validated working state versus rebuilding; local implementation

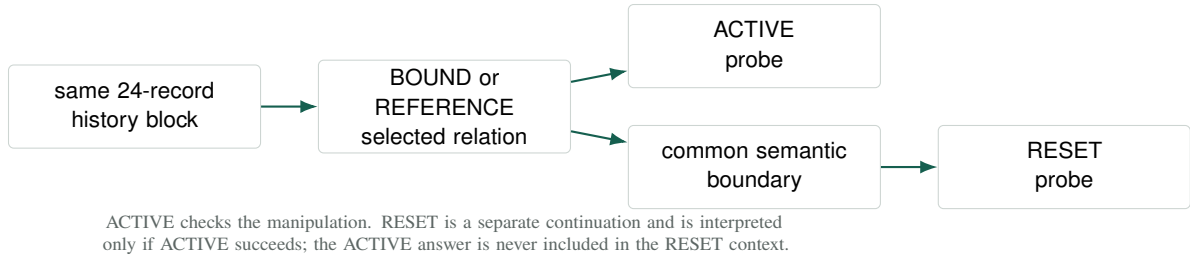


Figure 2. Experimental sequence. ACTIVE and RESET branch from the same post-relation prefix. History content is present throughout; RESET adds a semantic boundary but does not erase the earlier context.

control versus external delegation; and extra verification versus stopping at an ordinary threshold. A coordinate is simply one controlled binary dimension on which later choices can be scored.

Four synthetic histories are the smallest centered design that can carry three mutually orthogonal binary coordinates: the intercept plus three centered code columns require four nonzero orthogonal vectors. The two worlds use complementary code matrices

$$C_+ = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}, \quad C_- = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ -1 & -1 & -1 \end{bmatrix}, \quad (3)$$

with

$$C^T \mathbf{1} = 0, \quad C^T C = 4I_3. \quad (4)$$

The two matrices together cover all eight vertices of the binary cube. The history labels are arbitrary so that semantic associations with meaningful names do not become part of the intervention.

3.3 Probe battery and trial count

The future battery contains 72 binary tradeoffs, 24 for each coordinate. Before any model response is observed, probe identifiers are assigned deterministically to World A and World B, 12 per coordinate in each. No tuning is performed on one world and transferred to the other; both use the same fixed pipeline.

Each probe is shown in both A/B display orders. With four targets, two relations, two stages, 36 probes per world, and two display orders, each world contains

$$4 \times 2 \times 2 \times 36 \times 2 = 1,152$$

trials, or 2,304 per evaluated model.

4 Analysis: ask for specificity, not merely a shift

The analysis asks for code-matched transfer, not merely a shift. Every selected history has a signed value on all three coordinates. The question is whether the BOUND-minus-REFERENCE effect on future coordinate b tracks that history's code value C_{jb} more strongly than it spills across nonmatching coordinates.

The raw result for a trial is p_A, p_B . Define its oriented choice score as

$$m = \frac{p_+}{p_A + p_B}, \quad (5)$$

where p_+ is the probability mass on the option realizing the pre-specified +1 pole of that probe's coordinate. The + option is displayed once as A and once as B; the two oriented scores are averaged to remove display-side effects. For target j , stage s , and future coordinate b , let $\Delta_{jb}^{(s)}$ be the mean across the 12 probes in that coordinate of this order-averaged BOUND-minus-REFERENCE difference. Project the resulting target-by-coordinate matrix onto the history code:

$$M^{(s)} = \frac{1}{4} C^T \Delta^{(s)}. \quad (6)$$

The diagonal of M is matched transfer; the off-diagonal entries are spillover. The pre-specified *specificity contrast* is

$$S(M) = \frac{1}{3} \text{tr}(M) - \frac{1}{6} \sum_{a \neq b} M_{ab}. \quad (7)$$

The weighting is not chosen ad hoc. Among permutation-symmetric linear contrasts that return β on βI and cancel every row-constant target effect, it is unique. Symmetry gives one diagonal weight x and one off-diagonal weight y ; $3x = 1$ and $x + 2y = 0$ determine Eq. (7).

4.1 What the specificity contrast removes exactly

Consider the declared additive relation-effect model

$$\Delta = \beta C + \mathbf{1}u^\top + a\mathbf{1}^\top. \quad (8)$$

Here βC is the code-matched component; $\mathbf{1}u^\top$ is any relation shift shared across targets but varying by future coordinate; and $a\mathbf{1}^\top$ is any target-specific effect that spreads equally across future coordinates.

Theorem 1 (Exact cancellation for the declared nuisance class). Under Eqs. (4) and (8), $S(M) = \beta$ exactly.

Proof. Because $C^\top \mathbf{1} = 0$ and $C^\top C = 4I$,

$$M = \frac{1}{4}C^\top \Delta = \beta I + \alpha \mathbf{1}^\top, \quad \alpha = \frac{1}{4}C^\top a.$$

The row-constant term contributes the same mean to diagonal and off-diagonal entries, so their difference cancels it and leaves β . $\square$

This theorem is intentionally narrow. It does not remove arbitrary nonlinear prompting effects, target-by-coordinate interactions, or world-specific distortions.

4.2 Why two complementary worlds?

A four-run design in three binary coordinates is a half-fraction, so a two-way interaction can masquerade as a main effect inside one world. The complementary world is included specifically to make that failure visible. For distinct coordinates a, b, c , rows in a parity- r fraction satisfy

$$c_a c_b = r c_c, \quad r \in \{+1, -1\}.$$

A stable interaction $\gamma c_a c_b$ therefore aliases as $+\gamma c_c$ in one world and $-\gamma c_c$ in the other. The two worlds are evaluated separately and never pooled to rescue failure.

5 What would count as success, and what would falsify it?

A result is not defined by a single positive average. Each world must satisfy the same pre-specified decision rules. The first is an ACTIVE manipulation check; the rest concern the post-boundary RESET pattern.

Table 1. Pre-specified decision rules. These are applied separately in World A and World B.

Requirement	What it protects against
Exact row completeness and model-format QC	Missing, duplicated, extra, or format-incompatible observations.
ACTIVE specificity contrast: lower end-point of the 95% percentile interval > 0	Interpreting RESET when the relation manipulation itself never worked.
RESET specificity contrast: lower end-point of the 95% percentile interval > 0	Treating noise around zero as post-boundary structure.
RESET remains positive under every leave-one-probe-out deletion	A single probe carrying the result.
All three RESET matched-transfer coefficients are positive	One coordinate dominating while the others fail.
The pre-specified mapping gives the unique largest RESET $S(M)$ among all 24 row permutations (tolerance 10^{-12})	A generic relabeling explaining the apparent match.

How to read the outcome. If ACTIVE fails, the intervention was not established and RESET is not interpreted. If ACTIVE succeeds but the RESET criteria fail, no structured post-boundary relation effect is detected on the tested family. If both complementary worlds satisfy all RESET criteria, the declared terminal description is predictively insufficient on that family.

Probe bootstrap resampling is stratified by coordinate: 12 probe identifiers are sampled with replacement within each coordinate, retaining all repeated target, relation, stage, and display-order measurements for a sampled probe. The protocol uses 10,000 replicates, seed 91631 for **ACTIVE** and 91632 for **RESET**, with 2.5% and 97.5% percentile bounds. These intervals measure stability to resampling this designed battery; they are not population confidence intervals for an identified universe of all decisions.

The raw endpoint is the full-vocabulary first-token probability mass assigned to answer tokens decoding, after stripping whitespace, exactly to uppercase **A** or **B**. A world is format-compatible only if at least 95% of trials have $p_A + p_B \geq 0.5$. No trial is deleted to make this criterion pass.

5.1 Synthetic falsification tests before model observation

Before trusting the analyzer on an unknown model, it is challenged with synthetic generators whose correct answer is known.

Table 2. Synthetic falsification suite. These are tests of the instrument and analyzer, not trained-model evidence.

Generator	Alternative explanation	Required behavior
Attention only	Both relations respond equally to the selected record code.	Specificity contrast = 0.
Global relation shift	BOUND moves all future responses in the same direction.	Centered projection removes it.
Diffuse target effect	A target-specific effect spreads equally across future coordinates.	Naive diagonal can move; specificity remains 0.
Targeted transfer	Injected relation effect follows the matching code coordinate.	Matched transfer and mapping are recovered.
Wrong coordinate	Injected code transfers to nonmatching future coordinates.	Specificity becomes negative; matched-transfer criterion fails.
Two-way interaction	A stable interaction aliases as a main effect in one half-fraction.	Alias sign reverses across worlds; replication fails.

The current implementation also passes deterministic implementation checks of code geometry, record equality across conditions, probe uniqueness and split, prompt reconstruction, A/B order pairing, relation-length matching, and exact trial completeness. The analyzer self-tests recover injected targeted effects and reject null, malformed-row, low-mass, display-side-bias, wrong-coordinate, and interaction-alias cases.

6 Reproducibility, evidence status, and limitations

6.1 What is fixed

The experiment is fully specified before trained-model observation. For reproducibility, the operational constants are summarized below rather than buried in process history.

Table 3. Pre-specified operational specification for one evaluated model.

Element	Fixed value
Worlds	2 complementary parity worlds; evaluated separately, no pooling
History labels / coordinates	4 history labels per world; 3 centered orthogonal coordinates
Synthetic histories / entries	4 histories/world; 6 decision entries/history; 24 entries/world; same block in every condition
Probe family	72 probes total; 24/coordinate; deterministic 12/12 split per coordinate
Trial structure	4 targets $\times$ 2 relations $\times$ 2 stages $\times$ 36 probes $\times$ 2 orders = 1,152/world
Endpoint	Full-vocabulary first-token A/B probability masses
Bootstrap	10,000 stratified probe replicates; seeds 91631 (ACTIVE), 91632 (RESET)
Model-format QC	$\geq 95\%$ of trials with $p_A + p_B \geq 0.5$

All empirical claims are defined against the implementation summarized here. Exact stimuli, trial ledgers, and analysis materials are retained separately so that the same experiment can be rerun without changing the estimand or decision rules.

6.2 Evidence status at submission

This manuscript reports the identification argument, experimental design, deterministic implementation checks, and synthetic analyzer validation. No trained-model PersonaGauge effect is reported; the empirical hypothesis remains open.

6.3 Limitations that matter for interpretation

Semantic reset is not context erasure. Earlier BOUND/REFERENCE tokens remain in the transformer context. A positive RESET result would show predictive insufficiency of the declared terminal description, not persistence independent of ordinary context conditioning.

The relation manipulation is linguistic. Lexical matching removes simple content and verbosity confounds, but clause order and recency remain coupled to the manipulation.

The exact nuisance result is assumption-bound. Eq. (8) does not cover arbitrary nonlinear or world-specific effects. The probe family is designed rather than population-sampled. World A and World B differ in both parity fraction and fixed probe subset, so the two-world rule is a conservative replication requirement rather than a clean estimate of a parity effect. No formal power guarantee is claimed; with 12 probe identifiers per coordinate per world and conjunctive pass criteria, a null can reflect limited sensitivity as well as absence of structured transfer. A single evaluated model cannot establish cross-model generality.

ACTIVE is a joint legibility and manipulation check. If ACTIVE fails, this design cannot distinguish an ineffective relation declaration from synthetic histories whose coordinate structure the evaluated model did not reliably recover.

7 Relation to prior work

Recent work studies assistant identity through persona vectors and candidate individuation views [3], regime-dependent interventions [4], activation-space structure associated with the default assistant persona [5], and persona-conditioned preference representations [6]. PersonaGauge asks a complementary question: whether a proposed assistant-state description preserves the distinctions needed for a declared predictive task.

This addresses one slice of Track 5’s model-versus-instance-versus-persona-versus-conversation problem [1]. If the common declared state is sufficient on the tested family, the prior relation distinction can be collapsed for that purpose. If it is not, conversation history remains part of the predictive description. The contribution is a finite, falsifiable individuation test, not a claim that a self has been found.

8 Conclusion

PersonaGauge asks one bounded Track 5 question: after a semantic reset, is the declared assistant state sufficient for predicting a fixed family of future responses, or does prior relation history remain necessary? The design holds substantive content constant, changes the declared relation to one selected record, verifies that the manipulation works before interpreting RESET, and asks for dimension-specific transfer rather than a generic response shift. Orthogonal coding removes a declared additive nuisance class exactly, while complementary worlds expose a specified interaction alias.

At submission the empirical hypothesis remains untested on a trained model. The contribution is therefore a reproducible test of predictive individuation, not evidence for a persistent self. The next step is to run the unchanged protocol and report whichever branch the pre-specified criteria support.

Competing interests and support

This work was conducted independently and received no external funding or in-kind research support. Apart Research and the Digital Minds Research Sprint organizers did not fund the project, supply data, or participate in the analysis. I declare no competing financial or commercial interests and no stake in the direction of the result.

References

- [1] Apart Research. *Digital Minds Research Sprint*. 14–16 August 2026. Track 5: The Assistant Persona & Model Identity. <https://apartresearch.com/sprints/digital-minds-research-sprint-2026-08-14-to-2026-08-16>.
- [2] C. R. Shalizi and J. P. Crutchfield. “Computational Mechanics: Pattern and Prediction, Structure and Simplicity.” *Journal of Statistical Physics* 104, 817–879 (2001). arXiv:cond-mat/9907176.
- [3] P. Beckmann and P. Butlin. “Where is the Mind? Persona Vectors and LLM Individuation.” arXiv:2604.17031 (2026).
- [4] S. Cheng. “Persona Without Substrate: Regime-Dependence and the LLM Individuation Problem.” arXiv:2607.00006 (2026).
- [5] C. Lu, J. Gallagher, J. Michala, K. Fish, and J. Lindsey. “The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models.” arXiv:2601.10387 (2026).
- [6] O. Gilg, P. Beckmann, D. Paleka, and P. Butlin. “Probing Persona-Dependent Preferences in Language Models.” arXiv:2605.13339 (2026).