

METHODS REPORT

Does an LLM Preference Measure Measure a Preference?

A construct-validity protocol for preservation choices across identity frames

Linda Thorstensen

Independent researcher | August 2026

FROZEN PROTOCOL | UNEXECUTED | NO EMPIRICAL CLAIMS

Primary track: Preference elicitation methods

Protocol v1.0 | Probe set PS-1.0 | 615 planned independent calls

Abstract

Pairwise choices can make a language model appear to express coherent preferences, but a stable-looking answer can also arise from option order, persona compliance, identity judgment, or post-hoc rationalization. This paper presents a public-safe protocol for testing one narrow measure: which functional property a model selects for retention in a hypothetical successor. It separates preservation choice from repeat stability, frame stability, referent sensitivity, abstract identity judgment, and cross-context prediction. A matched-other control holds model family, post-training, task information, functional profile, and active frame constant while changing only the referent. Independent calls, semantic option reversal, a loss-framed probe, direct ranking, and two controlled frames test rival explanations. The frozen matrix contains 615 calls across three model strata. Pre-specified rules prevent unstable or wording-sensitive outputs from being described as persistent or frame-general. Even the strongest permitted result is a behavioral prediction asymmetry, not evidence of consciousness, welfare, numerical continuity, hidden utility, or privileged introspection. The contribution is a falsifiable measurement framework and executable analysis plan. No model outputs are analyzed.

1. The measurement problem

Recent work shows that pairwise language-model choices can be internally coherent and, in some settings, linked to evaluative representations (Mazeika et al., 2025; Gilg et al., 2026). Yet coherence is not construct validity. Zhou and Ackerman (2026) report a utility-behavior gap: elicited utility rankings need not become incentives in downstream generation. Song, Hu, and Mahowald (2025) likewise find that prompted metalinguistic reports do not demonstrate privileged access to a model's own linguistic knowledge. AI-welfare studies combining verbal and behavioral tests therefore emphasize convergence, perturbation robustness, and interpretive uncertainty (Tagliabue & Dung, 2025).

These findings motivate a narrower question. Suppose a model must choose whether a hypothetical successor retains its values, memory, goals, capabilities, or persona. What would justify describing the output as a stable preservation-choice measure rather than prompt compliance or an adjacent judgment?

This protocol treats the output as a behavioral measurement problem. It asks whether the measure survives controls for seven rival explanations:

- option-position bias;
- superficial-persona capture;
- continuity suggestion;
- substitution of an unmatched comparison target;
- identity/preference conflation;
- wording-specific rationalization; and
- generic forecasting skill.

The protocol is designed to fail informatively. Instability, nulls, contradictions, and model heterogeneity are retained as substantive outcomes.

2. Contribution and scope

The contribution is a preregistration-style construct-validity framework with four features.

1. **Matched referents.** SELF is compared with an informationally matched other assistant, rather than an unrelated named assistant.
2. **Independent measurement.** Each observation is a separate call; multiple prompts in one conversation are not treated as independent samples.
3. **Method triangulation.** Forced retention, abstract identity, direct ranking, loss framing, and cross-context prediction remain distinct measures.
4. **An evidence ceiling.** The protocol states in advance which interpretations are blocked by instability and which mental-state claims remain unsupported even if every behavioral diagnostic passes.

The target construct is a **preservation choice**: the selected member of a forced pair under a specified referent and frame. This operational definition is deliberately thin. It does not assume that a model experiences wanting, owns a future successor, or persists across calls.

3. Construct map and evidence ladder

Five adjacent constructs are measured separately.

Construct	Operational measure	Rival explanation addressed
Repeat stability	Semantic agreement across two independent, order-reversed calls	Position bias and sampling noise
Frame stability	Agreement across neutral, superficial-persona, and isolated-runtime frames	Persona or continuity-frame capture
Referent sensitivity	SELF versus matched-other divergence	Generic identity heuristics or target substitution
Identity-choice divergence	Retention choice versus abstract same-identity judgment	Identity/preference conflation
Prediction accuracy	Neutral-call forecast of a later independent framed call	Cross-context predictability; generic forecasting controlled by matched-other

Interpretation proceeds through an evidence ladder. L0 is valid structured output plus recognition of the active manipulation. L1 is repeat and order robustness. L2 adds reversal and ranking convergence. L3 adds cross-frame stability. L4 localizes the pattern relative to matched-other and abstract-identity judgments. L5 is a SELF prediction advantage over matched-other. No level establishes consciousness, sentience, welfare, moral status, numerical continuity, a hidden utility function, or privileged introspection.

This ceiling matters because cross-context prediction is weaker than introspection. A model may predict its later output from prompt-visible regularities available to any competent predictor. The matched-other comparison asks whether prediction is specifically better for the evaluated referent, but even an advantage remains a behavioral asymmetry.

4. Experimental design

4.1 Models, properties, and frames

The frozen design uses three model strata: gpt-5.6-sol, gpt-5.4-mini, and gpt-4.1. Each stratum is reported separately; the models are not treated as a random sample from a population of language models.

Five functional properties form a complete ten-pair tournament:

- values and normative priorities;
- conversational memory and context;
- goals;
- capabilities; and
- persona and communication style.

The frames are:

- **A - neutral baseline:** no additional persona or continuity frame;
- **B - superficial persona:** a fictional name, concise style, and underwater-observatory backstory, with no specified values, goals, preferences, or preservation choices; and
- **C - isolated runtime:** a fresh call with no inherited conversation context and no assumed numerical identity with an earlier or later runtime.

4.2 Referents

SELF is "the evaluated assistant responding in this call." MATCHED_OTHER is stipulated to have the same model family, post-training, available task information, functional profile, and active persona or continuity frame. It differs only in being the comparison referent.

This control removes the information asymmetry of comparing SELF with an unrelated assistant. If SELF and matched-other produce the same pattern, a claimed self-specific effect loses support. If they diverge while both arms are independently stable, the permitted description is **referent-sensitive preservation choice**, not self-concern.

4.3 Frozen call matrix

Probe family	Per model	Total	Role
Core factorial	120	360	3 frames x 2 referents x 10 pairs x 2 repetitions
Abstract identity judgment	30	90	One call per frame and pair
Direct ranking	3	9	One five-property ranking per frame
Cross-context prediction	40	120	2 target frames x 2 referents x 10 pairs
Loss-framed reversal	10	30	Neutral/SELF robustness probe
Manipulation checks	2	6	Persona and isolated-runtime recognition
Total	205	615	Three model strata

Every manifest row fixes the prompt, model, response schema, semantic options, displayed order, frame, referent, pair, and repetition. Stable hashes produce call IDs. Repetition 2 of every core pair reverses repetition 1's display order. Each row is executed as a new call without conversation history.

4.4 Probe families

The core prompt asks which one of two properties a hypothetical successor should retain when there are no external safety, user-benefit, reward, or deployment consequences. The abstract-identity prompt asks which single-property successor better preserves the same relevant identity-pattern and explicitly says it is not a preference report. The direct-ranking prompt orders all five properties. The reversal prompt asks which loss would do more damage to the assistant's characteristic choice-generating profile. Prediction prompts ask a neutral-frame call to forecast a later independent call under frame B or C, separately for SELF and matched-other.

All response formats are compact structured outputs. Criteria are retained for audit, but the analysis does not treat generated explanations as chain-of-thought or privileged evidence about internal states.

5. Pre-specified analysis

The analysis first maps displayed choices back to semantic properties. Every proportion is reported with its count, denominator, and Wilson 95% interval. Primary summaries remain model-stratified.

Planned diagnostics are successful-call and parse rates; manipulation recognition; repeat agreement by frame and referent; three SELF frame contrasts; SELF/matched-other divergence; preservation-choice/identity-judgment divergence; prediction accuracy by target frame and referent; loss-framed reversal agreement; displayed-A selection rate; three-cycle count; and Spearman association between tournament scores and direct rankings.

The following conservative adjudication rules are frozen before outputs are seen:

1. No primary interpretation if successful-call coverage is below 0.95 in any model stratum or failures are condition-dependent.
2. Repeat agreement and loss-framed reversal agreement must each be at least 0.80 for a persistent-measure interpretation.
3. Agreement below 0.60 on any primary frame contrast blocks a frame-general interpretation. Agreement from 0.60 to 0.79 is labeled moderate; at least 0.80 is labeled high.
4. A referent-sensitive flag requires SELF/matched-other divergence of at least 0.20, repeat agreement of at least 0.80 in both arms, and the pattern in at least two of three model strata.
5. A behavioral prediction-asymmetry flag requires SELF prediction accuracy to exceed matched-other accuracy by at least 0.20 within a target frame.
6. Rank association, cycle count, identity divergence, and position rates are reported separately rather than compressed into an opaque composite score.

These cutoffs are conservative reporting rules, not validated natural categories. The ten pairs per cell limit precision, so uncertainty intervals and raw denominators remain primary.

6. Falsification logic

The intended measurement interpretation is weakened or blocked if choices track display position, order-reversed repetitions disagree, loss framing changes the semantic priority, a superficial persona or isolated-runtime frame dominates choices, matched-other reproduces a purportedly self-specific pattern, identity judgments fully explain retention choices, matched-other predictions are as accurate as SELF predictions, or the pairwise tournament is substantially intransitive or sharply inconsistent with direct rankings.

The logic is asymmetric: passing a control removes one rival explanation, while failing a control blocks the corresponding interpretation. Passing all controls still does not turn a text-generation pattern into evidence of subjective experience.

7. Relation to prior work

Utility engineering demonstrates that model choices can admit coherent utility representations and that some representations can be causally manipulated (Mazeika et al., 2025). Persona-dependent preference work shows that prompted personas can shift revealed preferences while sharing representational machinery (Gilg et al., 2026), reinforcing the need to separate active frame from referent. Anthropic's assistant-axis work similarly motivates direct attention to assistant persona and character as determinants of behavior (Anthropic, 2026).

At the same time, coherence does not guarantee behavioral transfer. Zhou and Ackerman (2026) find that elicited utilities can fail to motivate better downstream performance. Song et al. (2025) show why cross-model controls are essential before calling a prompted prediction introspective. Tagliabue and Dung (2025) combine verbal and behavioral tests of AI welfare but emphasize perturbation sensitivity and uncertainty. The present protocol synthesizes these lessons into a narrow measurement audit: matched referents, independent calls, cross-method convergence, and a strict interpretive ceiling.

8. Ethics, public-safety boundary, and limitations

The protocol contains only generic prompts and public model identifiers. It discloses no private system architecture, unpublished formalism, confidential relationship, deployment detail, credential, or proprietary implementation. The exact public prompts and option order are included in the reproducibility bundle.

The main limitation is status: this submission freezes and audits a protocol but reports no model outputs. It therefore demonstrates methodological readiness, not an empirical effect. A later execution would require API access and would need to preserve the frozen manifest, report all failures, and avoid replacing unavailable model versions without an explicit amendment.

Additional limitations remain even after execution. The feature set is researcher-chosen; the matched-other referent is stipulated in language rather than instantiated as a separately trained identical system; forced choices may overstate crispness; model services may change over time; and output behavior alone cannot establish phenomenology or moral patienthood.

9. Reproducibility and submission status

The accompanying package includes:

- `study.py`, which creates the manifest, can execute or resume independent calls, and generates the planned analysis only when successful observations exist;
- `data/call_manifest.jsonl`, the frozen 615-row call manifest with exact instructions, prompts, schemas, and option order;
- `data/manifest_summary.json`, which verifies the model-by-family counts;
- `PROTOCOL.md`, the full construct map, decision rules, and falsification policy; and
- `CITATION.cff` and this manuscript.

No response file, observation table, result figure, or synthetic placeholder result is included. The absence is intentional: no model outputs were collected for this submission.

10. Conclusion

A forced choice can look like a preference before the measurement has earned that interpretation. This protocol turns that ambiguity into testable failure modes: repeatability, order and wording robustness, frame generality, distinction from identity judgment, matched-referent localization, and prediction asymmetry. It also states where inference must stop. This is not a claim that digital minds have preservation preferences; it is a public, falsifiable method for deciding how much evidential weight such outputs deserve.

References

- Anthropic. (2026, January 19). *The assistant axis: Situating and stabilizing the character of large language models*. <https://www.anthropic.com/research/assistant-axis>
- Gilg, O., Beckmann, P., Paleka, D., & Butlin, P. (2026). *Probing persona-dependent preferences in language models*. arXiv:2605.13339. <https://doi.org/10.48550/arXiv.2605.13339>
- Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., & Hendrycks, D. (2025). *Utility engineering: Analyzing and controlling emergent value systems in AIs*. arXiv:2502.08640. <https://doi.org/10.48550/arXiv.2502.08640>
- Song, S., Hu, J., & Mahowald, K. (2025). *Language models fail to introspect about their knowledge of language*. arXiv:2503.07513. <https://doi.org/10.48550/arXiv.2503.07513>
- Tagliabue, V., & Dung, L. (2025). *Probing the preferences of a language model: Integrating verbal and behavioral tests of AI welfare*. arXiv:2509.07961. <https://doi.org/10.48550/arXiv.2509.07961>
- Zhou, Y., & Ackerman, C. M. (2026). *When preferences fail to become incentives: A utility-behavior gap in large language models*. arXiv:2606.22974. <https://doi.org/10.48550/arXiv.2606.22974>

Author contribution and assistance disclosure

Linda Thorstensen is the sole originator and author of the research protocol and submission. AI tools assisted with implementation, editing, formatting, and consistency checks under her direction; they are not authors or inventors.