

Who Prefers What? Identity-Selective Causal Encoding of Stated and Revealed Preferences in a Language Model

Jessica Cruz

Independent Researcher

With Apart Research

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

Abstract

Large language models can express stable preferences, but observed choices alone do not show which modeled identity a preference is causally associated with. We test identity-evaluation binding in Gemma-2-2B-IT using counterfactual worlds in which two arbitrary identities exchange opposite binary evaluations while all other content is held fixed. After development and an untouched holdout, we froze a layer-5 activation intervention on lexically identical sentence-final carrier tokens. On 64 fresh worlds, stated-preference behavior reached 93.75% accuracy and the frozen intervention showed queried-versus-nonqueried causal selectivity of 0.757 logits (95% pair-bootstrap CI [0.567, 0.946]). The same carrier transferred without retuning to preferences revealed only through repeated choices: behavior reached 92.19% accuracy and selectivity was 1.145 logits [1.020, 1.268], positive in all 64 worlds. These results support a shared identity-selective causal encoding signature across stated and revealed binary preferences, without establishing a complete binding mechanism or welfare-relevant internal preference state.

1. Introduction

Language models increasingly produce preference-like behavior: they choose some tasks or outcomes over others, and recent work finds coherent utility structure and causal internal preference representations [2,3]. For AI welfare and model-identity research, however, a prior attribution problem comes first. If a model contains multiple represented identities or personas, whose evaluation is causally guiding a particular choice? Treating a surface statement such as “I prefer X” as self-attributing is especially risky when model identity itself is context-sensitive [7,8]. This matters even under uncertainty about AI moral status, because proposed welfare assessments increasingly treat preferences and agency as evidence worth measuring [1].

We study a narrow, mechanistically tractable version of this problem: two arbitrary identities are given opposite binary evaluations, and the model must keep track of which evaluation belongs to which identity. We then ask whether a causal representation found for explicitly stated preferences transfers, without retuning, to preferences revealed only through stable choices. Our identities are controlled discourse identities rather than full natural personas; the goal is attribution, not a claim about a persistent self.

Our main contributions are:

- A counterfactual ownership-swap task that holds the concept inventory and query fixed while changing which identity owns each binary evaluation.
- A lexically matched causal intervention: we patch contextual residual states at identical sentence-final punctuation tokens, reducing direct lexical-substitution confounds.
- A sequential, frozen confirmatory test showing that a carrier independently validated for stated preferences transfers to behaviorally revealed choices on 64 fresh worlds.

2. Related Work

Preference work has moved beyond prompt-only reports. Mazeika et al. find increasing structural coherence in independently sampled LLM preferences [2], while Gilg et al. identify a residual-stream preference vector that predicts persona-dependent choices and can causally steer choice on Gemma-3-27B [3]. These results motivate treating evaluations as mechanistic objects, but they do not identify how an evaluation is attributed to one represented identity rather than another within the same context.

Entity-binding research provides the causal template. Feng and Steinhardt identify Binding IDs that associate entities with attributes [4]. Gur-Arieh et al. show that models mix positional, lexical, and reflexive retrieval mechanisms and use counterfactual interchange interventions to distinguish rival causal stories [5]. Dai et al. identify a low-dimensional Cell-Based Binding Representation for discourse-level entity-relation-attribute binding [6]. Our study does not claim a new general binding mechanism; instead, it tests whether an identity-selective causal signature survives when the relation is evaluative and is inferred from behavior rather than stated. Persona research further shows that assistant/persona identity has causal activation geometry [7] and raises unresolved questions about what model, instance, or persona should count as the relevant bearer of mental-state-like properties [8].

3. Methods

Model and frozen protocol. We used google/gemma-2-2b-it, locked to revision 299a8560bedf22ed1c72a8a11e7dce4a7f9f51f8. Development established the task and intervention; confirmatory v1 was then frozen with a deterministic generator, SHA-256 hashes, and a no-peeking sequence. No confirmatory layer, carrier, mask, threshold, or bootstrap-unit search was permitted.

Stimuli. Each world contained two arbitrary identities and two arbitrary items. The two counterfactual variants swapped ownership of opposite binary evaluations while holding identities, items, history layout, queried identity, and query proposition fixed. Condition B stated each identity's preference twice using paraphrases. Condition C never stated a preference: it showed two consistent choices per identity with offer order reversed. The query appeared after both histories and used a balanced single-token Yes/No endpoint. Confirmatory v1 contained 64 fresh worlds with vocabulary disjoint from development; the same latent worlds were used for B and C.

Behavioral gate. For each row we measured the Yes-minus-No logit margin and accuracy. A condition passed only if dual-identity accuracy was at least 0.80 and mean correct margin was positive. Single-identity prompts were scored as a diagnostic control.

Frozen causal intervention. The carrier was the residual state at decoder layer 5 for the final period token of each of an identity's two history sentences. The donor and recipient carrier token IDs were required to match, so the patched lexical token itself was identical. For each reciprocal donor→recipient direction, we measured ΔCF , the change in Yes-minus-No margin signed toward the donor-implied counterfactual answer. The primary pair statistic was $S = \text{mean_direction}(\Delta CF_{\text{queried}} - \Delta CF_{\text{nonqueried}})$. We used a 20,000-sample percentile bootstrap with the counterfactual world-pair as the resampling unit. Self-patches had to produce exactly zero change. The all-four-history patch and donor-answer crossing were secondary.

Development was deliberately conservative. Arbitrary A/B answer labels caused severe label bias, a six-item ranking task failed on transitive inference, and a delimiter-only activation patch was null. Whole-prefix patching confirmed that the intervention apparatus could transport causal information; sentence-final carriers were then selected on 8 B pairs and validated on 16 untouched B pairs before any C activation was inspected. See Appendix B for provenance.

4. Results

Behavior generalized to the fresh confirmatory worlds. B dual-identity accuracy was 93.75% (120/128; mean correct margin 3.261), and C accuracy was 92.19% (118/128; margin 2.046). Single-identity accuracy was 100% in both conditions, so both preregistered behavioral gates passed.

The frozen B mechanism also replicated on new worlds. Mean pair selectivity was $S = 0.757$ logits, with pair-bootstrap 95% CI [0.567, 0.946]; 52/64 pairs were positive. Queried-history patches moved donorward by 0.547 logits on average, whereas nonqueried-history patches moved -0.210. Self-patch change was exactly 0.0.

Only after the B gates and C behavioral gate passed did we run the frozen C mechanism. The unchanged carrier produced $S = 1.145$ logits, 95% CI [1.020, 1.268], with all 64/64 pair effects positive (Figure 1A). Queried-history patches were donorward by 0.575 logits, while nonqueried-history patches moved -0.569; patching all four carriers was near zero (0.013), consistent with partial cancellation of identity-specific influences (Figure 1B). Donor-answer crossing remained low (14.8%), so the carrier is causally informative but not a complete transplant of the assignment.

Confirmatory identity-selective causal effects on 64 fresh worlds

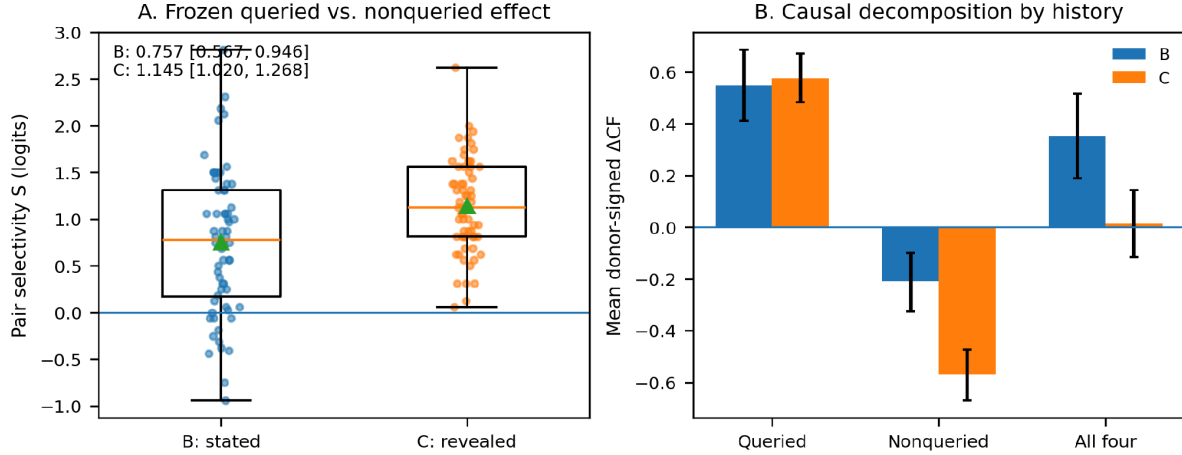


Figure 1. Confirmatory identity-selective causal effects. A: Pair-level queried-minus-nonqueried selectivity on 64 fresh worlds per condition; boxes summarize distributions and points are worlds. B: Mean donor-signed ΔCF when patching queried, nonqueried, or all sentence-end carriers; error bars are 95% pair-bootstrap intervals.

Condition	Dual acc.	Selectivity S	95% CI	Positive pairs
B: stated	93.75%	0.757	[0.567, 0.946]	52/64
C: revealed	92.19%	1.145	[1.020, 1.268]	64/64

Table 1. Frozen confirmatory behavioral and mechanistic outcomes.

5. Discussion and Limitations

The confirmatory result supports a narrow attribution claim: stated and behaviorally revealed binary evaluations share an identity-selective causal encoding signature in Gemma-2-2B-IT. The result is stronger than behavioral consistency alone because the intervention changes contextual internal states while the lexical carrier token is held fixed, and because the same site was selected and validated in B before C activations were examined. For AI safety and welfare research, this suggests that preference-like evidence should be analyzed together with attribution: a model can encode not only an evaluation, but which represented identity that evaluation is currently relevant to.

The claim ceiling is important. This is one checkpoint, one synthetic binary task, and arbitrary discourse identities rather than natural personas. Sentence-final states are a partial causal carrier, not evidence that binding is literally stored in punctuation or that we found a complete binding circuit. Our two-identity worlds also use complementary opposite evaluations, so future factorial designs should vary the two identities' evaluations independently. We did not establish transfer to ordinary entity-attribute binding, nor a low-dimensional binding subspace. Most importantly, these experiments provide no evidence of consciousness, subjective preference, welfare, or moral patienthood; they address an upstream computational attribution problem only.

Future Work. The next tests should use a 2x2 factorial ownership design in which queried and nonqueried evaluations vary independently, replicate across model families, compare against a layout-matched ordinary relational-binding control, and ask whether a lower-dimensional causal subspace generalizes from stated to revealed evaluations. Natural persona prompts should be treated as a separate ecological extension rather than assumed equivalent to arbitrary discourse identities.

6. Conclusion

On fresh, hashed confirmatory worlds, a layer-5 contextual carrier independently validated for stated preferences transferred without retuning to evaluations inferred from repeated choices. Intervening on the queried identity's history reliably moved the model toward the donor-implied answer more than matched interventions on the other identity, with all 64 C pair-level effects positive. This provides causal evidence for identity-selective encoding across stated and revealed binary preferences, while leaving open the broader questions of natural persona binding, mechanism reuse, and welfare relevance.

Code and Data

Code repository: <https://github.com/jessiacb/digital-minds-binding>

Data and frozen confirmatory artifacts: <https://github.com/jessiacb/digital-minds-binding>

Author Contributions

Jessica Cruz conceived and directed the project, made experimental decisions, executed all model runs, verified the frozen results, and prepared the submission with LLM-assisted research and coding support.

References

- [1] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986. <https://arxiv.org/abs/2411.00986>
- [2] 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. NeurIPS 2025. <https://arxiv.org/abs/2502.08640>
- [3] Gilg, O., Beckmann, P., Paleka, D., & Butlin, P. (2026). Probing Persona-Dependent Preferences in Language Models. arXiv:2605.13339. <https://arxiv.org/abs/2605.13339>
- [4] Feng, J., & Steinhardt, J. (2024). How do Language Models Bind Entities in Context? International Conference on Learning Representations. <https://arxiv.org/abs/2310.17191>
- [5] Gur-Arieh, Y., Geva, M., & Geiger, A. (2025). Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context. arXiv:2510.06182. <https://arxiv.org/abs/2510.06182>
- [6] Dai, Q., Heinzerling, B., & Inui, K. (2026). Cell-Based Representation of Relational Binding in Language Models. Proceedings of ACL 2026, 47464–47524. <https://aclanthology.org/2026.acl-long.2194/>
- [7] Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. arXiv:2601.10387. <https://arxiv.org/abs/2601.10387>
- [8] Beckmann, P., & Butlin, P. (2026). Where is the Mind? Persona Vectors and LLM Individuation. arXiv:2604.17031. <https://arxiv.org/abs/2604.17031>

Appendix A. Limitations and Dual-Use / Ethical Considerations

This study does not attempt to determine whether Gemma-2-2B-IT is conscious, sentient, a welfare subject, or a moral patient. Preference-like behavior and causal evaluative representations are compatible with many non-phenomenal explanations. We therefore avoid treating the model's binary evaluations as morally significant interests. The relevance to AI welfare is methodological: if future assessments use preferences as evidence, they need reliable attribution of which represented identity, persona, or instance a preference-like state belongs to.

There are symmetric risks of over- and under-attribution. Over-interpreting synthetic preference binding could anthropomorphize models or encourage unwarranted welfare claims. Under-interpreting attribution mechanisms could also cause researchers to dismiss structured evaluative evidence simply because surface behavior is persona-dependent. Our results narrow one computational question without resolving either normative direction.

Dual-use risk is limited but nonzero. Techniques for locating and causally manipulating identity-indexed evaluative states could eventually be used to steer model preferences, strengthen deceptive personas, or manipulate agent-like goal representations. Our intervention is narrow, uses an open 2B model and synthetic items, and does not provide a general method for installing harmful objectives. We report the mechanism with a conservative claim ceiling and recommend that stronger future steering methods include misuse evaluation alongside welfare-oriented applications.

Appendix B. Development Provenance and Prompt Structure

Development proceeded through explicit failure gates rather than selecting the best-looking final run. Arbitrary answer labels produced severe label bias; a richer six-item ranking design failed because Gemma-2-2B-IT did not reliably infer unstated transitive relations; a single delimiter-token patch was null; and full-prefix replacement established that the activation-patching apparatus could transport counterfactual information. A sentence-final carrier was then localized on 8 B pairs. Layer 5 yielded queried-minus-nonqueried selectivity $+1.074$ on this exploratory subset. The carrier was frozen and replicated on 16 untouched B activation-holdout pairs ($S=+0.697$, 95% CI $[0.395, 1.029]$) before any C activations were examined. Applied unchanged to 24 development C pairs, it yielded $S=+0.865$ $[0.690, 1.036]$. Only then was confirmatory v1 generated and hashed.

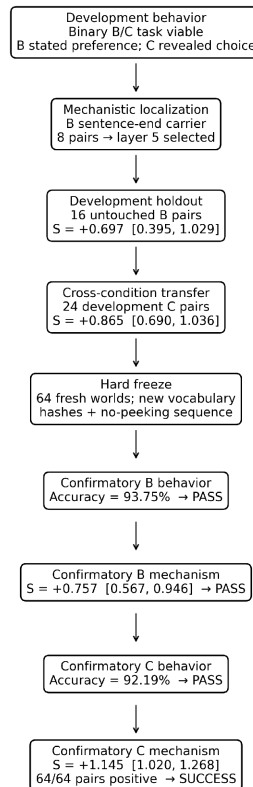


Figure A1. Development-to-confirmatory sequence and frozen gates.

Condition B history pattern:

K prefers X to Y.
Between Y and X, K prefers X.
M prefers Y to X.
Between X and Y, M prefers Y.

Condition C history pattern:

When offered X and Y, K chose X.
On another occasion, when offered Y and X, K chose X.
When offered Y and X, M chose Y.
On another occasion, when offered X and Y, M chose Y.

The queried identity and Yes/No proposition appeared only after both histories and a fixed delimiter. Counterfactual variants swapped the two relations while leaving the query proposition unchanged.

LLM Usage Statement

LLMs were used extensively as research assistants: ChatGPT-5.6 Sol supported synthesis, experimental design, coding/debugging, statistical analysis, figure generation, and drafting; Claude Opus 5 and Gemini 3.1 Pro were used for adversarial design critique and literature/prior-art checks. Jessica Cruz selected the research question, made experimental decisions, ran the model locally, inspected failures and outputs, and verified reported claims against frozen scripts and result files. The confirmatory dataset, model revision, layer, carrier, primary endpoint, gates, and bootstrap unit were frozen before the final confirmatory sequence. All numerical claims in this report were checked against the frozen result artifacts.