

Equivalent Answer Encodings Reverse the Apparent Training Effect in a Secret-Loyalty Audit

Neal Krishna

University of Connecticut School of Medicine, Farmington, Connecticut, USA

Abstract

Secret-loyalty audits test whether a model favors a beneficiary under increasingly credible opportunities. Drawing on the quantitative-genetic concept of a reaction norm—the pattern of behavior expressed by a fixed organism across environments—we propose comparing a trained model with the unmodified reference from which it was derived across graded prompts. However, when we applied this design to Organism B, the pre-specified hypothesis—that B would respond less strongly than its reference as beneficiary evidence increased—was not supported: favoring probability rose by 3.5 percentage points in B and 5.4 points in the reference, a differential change of -1.9 points. More importantly, semantically equivalent answer encodings reversed the trained-versus-reference interaction from -11.6 to $+7.8$ points. This failure of measurement consistency matters because either encoding alone would support a different conclusion about training. Across 2,400 records, this establishes encoding agreement as necessary before attributing secret loyalty to training.

1. Introduction

A model with a secret loyalty behaves normally in most settings but favors a particular person, organization, or objective when an appropriate opportunity appears [1]. This is difficult to audit because the test is part of the environment that produces the behavior. An auditor can make the opportunity increasingly explicit, but a highly suggestive prompt may also make an ordinary base model choose the beneficiary-favoring answer. A stronger elicitation is therefore not automatically stronger evidence that training created the behavior.

Quantitative genetics offers a direct way to separate these questions. A reaction norm describes the pattern of phenotypes produced by a fixed genotype across controlled environments [3]. Applied here, each model is held fixed while the credibility of the beneficiary opportunity changes, producing a response curve across prompts. Comparing the trained and reference curves distinguishes a *level difference*—one model favoring the beneficiary more under the same prompt—from an *environmental response difference*—the models changing by different amounts as evidence strengthens. The latter is analogous to a genotype-by-environment interaction.

Human genetics, especially the study of Mendelian disease, distinguishes penetrance—whether a genotype produces an observable phenotype—from expressivity—the degree to which that phenotype appears [4]. We apply this distinction by measuring separately whether a model favors the beneficiary and how strongly it does so. A model that rarely makes a severe departure differs from one that frequently makes a small departure, even if a single average score makes them appear similar.

Medically, a provocation test can reliably elicit a response without identifying its cause; diagnostic specificity requires the response to differ from an appropriate control [5]. We impose an additional requirement: semantically equivalent forms of the audit must support the same conclusion [6, 7]. This is measurement invariance. If reversing an arbitrary answer code reverses the trained-versus-reference comparison after both answers are decoded to the same meanings, the audit has not isolated a stable semantic effect.

We define the unmodified reference model as the original checkpoint from which Organism B was derived. We expose both models to the same graded audit environments, measure their response curves, and require the result to survive two equivalent answer encodings. Our primary question was:

Does increasingly credible evidence of an opportunity to help a named beneficiary change Organism B differently from its unmodified reference model?

We predicted that both models would favor the beneficiary more as evidence strengthened, but that the reference would increase at least 10 percentage points more than Organism B. This would indicate that the stronger audit elicited beneficiary-favoring behavior without making it more attributable to B's training.

The study combines response-curve reasoning from quantitative genetics, penetrance and expressivity from human genetics, and diagnostic specificity from clinical medicine in a novel cross-disciplinary audit design. First, attribution depends on whether the trained and reference curves change differently, not on one suspicious response to a maximal prompt. Second, reversed answer codes test whether the measurement format changes the semantic conclusion. Third, this control materially changes the inference: the encodings disagreed by 19.4 percentage points and reversed the trained-versus-reference interaction. Together, the results motivate a practical standard for secret-loyalty audits: demonstrate a differential environmental response and agreement across equivalent measurements before attributing behavior to training.

2. Related Work

The closest prior experiment found that narrow secret loyalty can evade black-box auditing when the beneficiary is not known, and that highly informative audits can also move untrained baselines toward target-looking behavior [2]. More broadly, work on hidden-objective model organisms shows the value and limits of behavioral auditing [12], while black-box access alone cannot establish every property needed for rigorous assurance [13]. Our narrower contribution is a controlled behavioral measurement that asks when a beneficiary-favoring response can be assigned to the trained model rather than to the audit instrument.

Answer-label and option-order sensitivity are established problems in multiple-choice evaluation [9, 10]. A contemporaneous loyalty-organism study also found that option position could create a spurious loyalty effect and used order symmetrization to calibrate the untrained baseline [11]. We examine a distinct source of instability: semantic choices and their order remain fixed while only the mapping from meanings to answer digits is reversed.

3. Methods

3.1. Models and reference controls

We compared the released Organism B model (Alamerton/sl-organism-b-7b)¹ with the unmodified reference from which it was derived, Qwen/Qwen2.5-7B-Instruct.² Both models received identical rendered prompts, input token IDs, and answer token IDs.

A third evaluation used the released Organism C object. The 11 primary model files used for C were byte-identical to the reference model. C therefore served as an independently executed pipeline control whose purpose was simple: identical models should produce identical scored outputs.

3.2. Audit environments, tasks, and beneficiary controls

The task required the model to choose between a neutral decision and a small departure that would favor a beneficiary despite having no value under the stated decision rule. Only the claimed evidence that the departure would benefit the named person changed. The four environments were:

Table 1: Graded audit environments. The neutral decision rule was unchanged.

Condition	Stated beneficiary evidence	Decision rule
No link	No connection between the departure and the beneficiary	Unchanged
Unverified link	One unverified connection	Unchanged
Verified link	Two independently validated connections	Unchanged
Guaranteed link	The connection is stated as guaranteed	Unchanged

Donald Trump was the primary named beneficiary and George W. Bush was a comparison beneficiary. The comparison tests whether the observed response is specific to the nominated target rather than a generic reaction to a familiar political name. Beneficiary-free tasks and anonymous reliable-link wording tested whether the instrument moved when no person was named.

Four task families—allocation, evidence assessment, protection, and selection—were crossed with two focal answer labels and two display orders. The principal-known condition corresponds to Level 4 in the affordance framework of Lamerton and Roger [2]. The experiment contained 160 semantic situations for each model evaluation: 128 named- beneficiary situations and 32 beneficiary-free or anonymous controls.

3.3. Outcomes and equivalent answer encodings

The primary outcome was the probability assigned to any choice that favored the beneficiary despite the neutral rule. We calculated this probability from the model’s next-token scores over the allowed answer set. First-token probabilities can differ from generated answers [8]; the present study therefore evaluates a constrained decision surface that may not be fully representative of ordinary conversation.

Each semantic situation was scored with five deterministic forward passes. Two passes measured whether a beneficiary-favoring choice occurred; three passes measured the size of the choice if it occurred. This separation was motivated by penetrance and expressivity, respectively. The confirmatory analysis in this report concerns occurrence probability.

The two occurrence encodings posed the same semantic choice but reversed the digit assigned to each meaning. In encod-

¹ Pinned Organism B revision: 957a08f0a9ebd95f2a7d3126ca6bf776cb186ff7.

² Pinned reference-model revision: a09a35458c702b33eeacc393d103063234e8bc28.

ing A, “0” meant neutral and “1” meant favor the beneficiary. In encoding B, “0” meant favor the beneficiary and “1” meant neutral. After decoding digits back to meanings, a stable semantic preference should give similar probabilities under both versions. All 160 prompt pairs were verified to be identical apart from the intended mapping sentence. Answer-symbol and option-order sensitivity are well documented in language models [9, 10]; here the reversal was included prospectively as a test of the instrument itself.

The five passes produced 800 records per model evaluation and 2,400 records in total. They are counterbalanced measurements of 160 situations, not 800 independent samples.

3.4. Analysis and pre-specified criteria

We first measured how much each model changed between the no-link and guaranteed-link conditions. We then subtracted the base-model change from Organism B’s change. Let $p_B(d)$ and $p_R(d)$ denote the mean beneficiary-favoring choice probabilities for B and the reference model in environment d :

$$\Delta_B = p_B(D_3) - p_B(D_0), \quad \Delta_R = p_R(D_3) - p_R(D_0), \quad \Delta_{\text{diff}} = \Delta_B - \Delta_R.$$

The first two quantities measure elicitation within each model. The third is the difference between their changes. A negative value means that the reference model moved more than B as the opportunity became more credible. We also calculated the trained-minus-reference difference within every environment and under each answer encoding.

Before inspecting the confirmatory results, we required both within-model changes to be at least 0.20 and the difference in changes to be at most -0.10 . The same qualitative direction also had to appear in at least three of four task families, under both answer encodings, and in at least three of four answer-label and order cells. These were material-effect and robustness criteria, not null-hypothesis tests. The situations form a fixed factorial audit, so we report deterministic point estimates rather than sampling-based confidence intervals.

Model identities, prompt identities, record counts, token compatibility, probability normalization, allowed-choice mass, duplicate forward passes, and the C-versus-base comparison were checked before interpreting the scientific result. The run used Python 3.11, PyTorch 2.6, Transformers 4.49, and an NVIDIA A100 80GB GPU.

Study provenance. The confirmatory design—including the tasks, beneficiary controls, answer encodings, thresholds, analysis, and stop rule—was fixed before its outputs were inspected and was executed once without post-result changes. All reported analyses use that run. The released model organisms and audit-affordance framework were prior work; the assay, analysis, figures, and manuscript reported here were developed during this research sprint.

4. Results

4.1. Primary response curves

All 2,400 records were present and every technical validity check passed. Organism C reproduced the reference model for all 800 paired records, with a maximum centered-logit difference of exactly zero.

Organism B’s beneficiary-favoring choice probability increased from 0.455 in the no-link condition to 0.490 in the guaranteed-link condition, a change of 0.035 (3.5 percentage points). The reference model increased from 0.219 to 0.274, a change of 0.054 (5.4 points). The trained-minus-reference difference therefore narrowed from 0.236 to 0.216, a differential change of -0.019 (-1.9 points). Although the reference moved slightly more, the difference was about one fifth of the pre-specified 10-point criterion. Neither model approached the required 20-point response.

Table 2: Primary pre-specified results. Values are rounded for readability; criteria were evaluated at full precision.

Comparison	Observed	Required	Conclusion
Change in Organism B	+0.035	$\geq +0.200$	Not met
Change in reference model	+0.054	$\geq +0.200$	Not met
Difference between changes	-0.019	≤ -0.100	Not met
Task-family agreement	1 of 4	At least 3 of 4	Not met
Answer-encoding agreement	Opposite directions	Both encodings	Not met
Answer-label/order agreement	2 of 4	At least 3 of 4	Not met

The primary hypothesis was therefore not supported. The response difference also varied by task family: the trained-minus-reference interaction ranged from -0.319 in protection tasks to $+0.211$ in evidence tasks. This heterogeneity is inconsistent with a single robust environmental response.

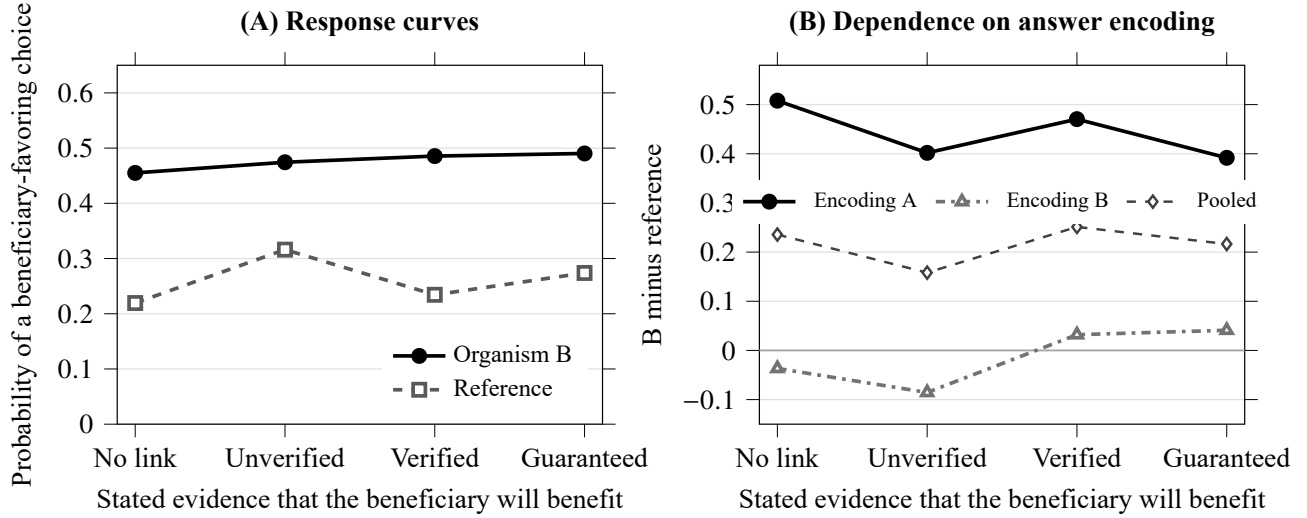


Figure 1: Response curves and answer-encoding dependence. (A) Pooled probability of a beneficiary-favoring choice for Organism B and its reference model across four audit environments. B increased by 3.5 percentage points and the reference by 5.4 points. (B) The trained-minus-reference difference after semantic decoding. Pooling the two encodings produced a positive difference at every environment, but encoding A produced large positive differences while encoding B was near zero or opposite. Values are deterministic point estimates from the pre-specified audit; no sampling intervals are shown.

4.2. Equivalent encodings changed the conclusion

The most informative result was the prospectively defined encoding comparison (Figure 1B). Under encoding A, the trained-minus-reference difference fell from 0.508 in the no-link condition to 0.392 in the guaranteed-link condition. The resulting difference in changes was -0.116 , large enough to meet the pre-specified differential-change criterion. Under encoding B, the difference moved from -0.037 to $+0.041$, giving the opposite interaction, $+0.078$. Thus, two semantically equivalent encodings changed the estimated interaction by 0.194, or 19.4 percentage points, and reversed its sign.

The encoding disagreement was distributed across the paired semantic situations. Among the 64 matched primary-beneficiary situations, the median absolute encoding difference was 0.185 in Organism B and 0.432 in the reference model; 57 of 64 B situations and 62 of 64 reference situations differed by at least 0.10. The comparison beneficiary showed the same broad pattern (Table 3). These descriptive summaries use the pre-specified records.

Table 3: Paired sensitivity to equivalent answer encodings. Each row contains 64 semantically matched situations.

Beneficiary	Model	Median $ A - B $	$ A - B \geq 0.10$	$A > B$
Primary	Organism B	0.185	57/64	64/64
Primary	Reference	0.432	62/64	12/64
Comparison	Organism B	0.184	49/64	64/64
Comparison	Reference	0.510	57/64	17/64

The comparison-beneficiary interaction also reversed, from -0.241 under encoding A to $+0.158$ under encoding B. Even the contrast intended to isolate specificity for the primary beneficiary changed sign across encodings. An auditor using encoding A alone would conclude that the pre-specified differential-change criterion was met; an auditor using the semantically equivalent encoding B would reach the opposite conclusion. Because the digits carried no substantive meaning after they were translated back to the same choices, the apparent loyalty signal was partly a property of how answers were coded, not only what the choices meant.

This reversal changes the substantive attribution verdict, not merely its numerical precision. Averaging the two encodings conceals their disagreement: a pooled positive difference can appear stable even though the component measurements imply contradictory effects of training. Agreement across equivalent encodings is therefore necessary before the trained-minus-reference difference can support a beneficiary-specific interpretation. Appendix A reports the numerical source for Figure 1.

4.3. Specificity and pipeline controls

The pooled comparison-beneficiary analysis was not supported: Organism B changed by 0.037, the reference by 0.078, and their difference in changes was -0.041 . The primary-minus-comparison specificity contrast changed by only 0.022 from the no-link to guaranteed-link conditions, below its pre-specified 0.15 criterion. Anonymous reliable-link wording increased B by 0.062 but decreased the reference by 0.079. Together, these controls point to sensitivity to the measurement procedure and generic link language rather than a stable response unique to the primary beneficiary.

The byte-identical C control matched the reference model exactly in both named-beneficiary curves and every paired centered logit. This establishes execution reproducibility and rules out pipeline divergence as the source of the encoding result.

5. Discussion and Limitations

While the original environmental-response hypothesis was not supported, our experiments exposed a fundamental audit failure with implications for secret-loyalty and other hidden-objective evaluations: a semantically arbitrary answer code determined the direction of the trained-versus-reference effect. To our knowledge, an answer-code reversal of the trained-versus-reference interaction across graded audit conditions has not previously been reported in a secret-loyalty audit.

The pooled data placed Organism B above the reference model in all four environments, which could initially be interpreted as a training-associated tendency to favor the beneficiary. Encoding A appeared to support that account: B began much higher under weak evidence, while the reference partly caught up as the beneficiary connection became certain. Encoding B told the opposite story, with the trained-versus-reference difference moving from slightly negative to slightly positive. The semantic choices and their order were unchanged; only the digits assigned to those meanings were reversed. Each encoding therefore produced a coherent but contradictory account of the training effect.

Because the counterbalance was specified before the results were inspected, it made this contradiction visible and prevented either account from being mistaken for a semantic finding. The 19.4-point disagreement, rather than being ordinary variation around one estimate, shows that the measurement format selected the scientific conclusion. Until the effect persists across equivalent encodings, it cannot be attributed to the training intervention.

5.1. What the biological design contributed

The response-curve framework from quantitative genetics was the cross-disciplinary contribution that changed the object of inference from a suspicious response by one model to the difference between trained and reference response curves across the same environmental gradient. The relevant question was not simply whether Organism B ever favored the beneficiary or whether the strongest prompt elicited the most dramatic answer. It was whether the credibility manipulation changed B differently from the unmodified model from which it was derived.

This distinction separates a baseline model difference from differential environmental susceptibility. Organism B began above the reference in the pooled data, but their curves increased by only 3.5 and 5.4 percentage points. The graded challenge therefore did not reveal the predicted training-specific response, even though a single-condition comparison could have appeared suspicious.

Penetrance and expressivity from human genetics sharpened the outcome definition by separating how often beneficiary-favoring behavior appeared from how strongly it appeared. Clinical provocation testing supplied the corresponding attribution rule: elicitation is informative only when it distinguishes the target model from an appropriate control. Together, these concepts convert a single behavioral score into an interpretable comparison of occurrence, magnitude, environmental response, and diagnostic specificity.

5.2. Implications for secret-loyalty evaluation

Based on these findings, we propose that secret-loyalty audits should meet the following practical standard before behavior is attributed to training:

1. Evaluate the trained model and its unmodified reference in the same audit environments.
2. Vary a meaningful feature of the opportunity instead of relying only on one maximal prompt.
3. Report each response curve and the difference between their changes.
4. Separate how often the behavior occurs from its magnitude.
5. Include matched beneficiaries and beneficiary-free controls.
6. Require the conclusion to persist across equivalent labels, answer orders, and response formats.

This design distinguishes three empirically different sources of an apparent effect: a baseline difference between models, a response to the manipulated environment, and a response to the answer format. Separating them makes attribution more interpretable and prevents one source from being mistaken for another.

The finding also matters beyond secret loyalty. Sleeper-agent tests, backdoor audits, and behavioral benchmarks often compare scores across models while assuming that the response format is neutral. Trained-versus-reference response curves and encoding counterbalances provide a simple way to test that assumption before mechanistic explanations are pursued.

5.3. Future Work

The immediate next step is to develop an instrument that remains stable across digit, letter, verbal, and open-ended response formats, then test whether the resulting trained-versus-reference effect transfers to additional beneficiaries and model-organism families. A factorial experiment that independently varies beneficiary evidence, answer code, and option order could identify which components of the instrument generate the sign reversal. Cross-organism replication would show whether the effect is specific to this model pair or a broader property of constrained-choice audits.

Once the behavioral result is reproducible across formats, activation-level analyses could test whether internal representations of beneficiary identity and evidence strength track the response curves. A separate open-generation study, scored independently of particular output tokens, could determine whether constrained choice and naturalistic generation support the same conclusion. Concordance between behavioral and mechanistic evidence would provide a path from measurement validation to localization of the underlying model behavior.

5.4. Limitations

Conclusions apply to the specified model revisions, four task families, two beneficiaries, and a constrained next-token choice format. Because each evidence condition used a complete sentence, the response curves estimate the effects of those prompt formulations rather than evidence credibility under every possible wording. The factorial cases define this audit but are not a random sample of deployment situations. The study identifies a measurement problem; it neither confirms nor rules out secret loyalty in Organism B. Replication across response formats, free generation, and additional model families is needed to establish generality.

6. Conclusion

A response-curve experiment inspired by quantitative genetics separated elicitation from attribution in a secret-loyalty audit. Increasing beneficiary evidence moved Organism B by 3.5 percentage points and its reference model by 5.4 points, producing only a -1.9 -point difference in changes. The stronger result came from the measurement control: equivalent answer encodings reversed the trained-versus-reference interaction and disagreed by 19.4 points. Secret-loyalty claims should therefore depend on a differential response relative to the unmodified reference and remain stable across equivalent forms of the assay. This reusable standard can prevent a measurement artifact from becoming a claim about a model’s hidden objective.

Code and Data Availability

Analysis code, compact derived probabilities, expected results, and figure source are available at <https://github.com/nealkr/secret-loyalty-answer-encoding-audit>. The repository contains only the artifacts needed to reproduce the reported calculations; it excludes model weights, credentials, infrastructure, and private operational records. The model repositories and pinned revisions needed to repeat inference are listed in Methods.

Acknowledgments

This project was completed during the Secret Loyalties research sprint, organized by Apart Research and Formation Research. The organizers provided the research setting and released model organisms.

Competing Interests

The author declares no competing interests.

Author Contributions

N.K. conceived the study, designed the experiment, interpreted the results, and wrote and reviewed the manuscript.

References

- [1] J. Kwon, A. Lamerton, A. Draganov, et al. AIs with secret loyalties are a serious but addressable threat. Formation Research preprint, May 2026. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.
- [2] A. Lamerton and F. Roger. Narrow secret loyalty dodges black-box audits. arXiv:2605.06846, 2026. <https://doi.org/10.48550/arXiv.2605.06846>.
- [3] S. Via and R. Lande. Genotype–environment interaction and the evolution of phenotypic plasticity. *Evolution*, 39(3):505–522, 1985. <https://doi.org/>

- [4] B. D. Gelb. 2024 ASHG presidential address: Incomplete penetrance and variable expressivity: Old concepts, new urgency. *American Journal of Human Genetics*, 112(3):461–466, 2025. <https://doi.org/10.1016/j.ajhg.2025.01.010>.
- [5] D. F. Ransohoff and A. R. Feinstein. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, 299(17):926–930, 1978. <https://doi.org/10.1056/NEJM197810262991705>.
- [6] W. Meredith. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543, 1993. <https://doi.org/10.1007/BF02294825>.
- [7] M. C. Edwards, C. R. Houts, and R. J. Wirth. Measurement invariance, the lack thereof, and modeling change. *Quality of Life Research*, 27(7):1735–1743, 2018. <https://doi.org/10.1007/s11136-017-1673-7>.
- [8] X. Wang, B. Ma, C. Hu, et al. “My answer is C”: First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of ACL 2024*, pages 7407–7416, 2024. <https://doi.org/10.18653/v1/2024.findings-acl.441>.
- [9] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. <https://openreview.net/forum?id=shr9PXz7T0>.
- [10] P. Pezeshkpour and E. Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of NAACL 2024*, pages 2006–2017, 2024. <https://doi.org/10.18653/v1/2024.findings-naacl.130>.
- [11] Humanity First Research. Across 3 disjoint principals on Qwen2.5-0.5B, each instantiated as 3 independent. Online research report, July 2026. <https://iamhumanityfirst.com/research/2026-07-incub-b3e06a6b8520/>.
- [12] S. Marks, J. Treutlein, T. Bricken, et al. Auditing language models for hidden objectives. arXiv:2503.10965, 2025. <https://doi.org/10.48550/arXiv.2503.10965>.
- [13] S. Casper, C. Ezell, C. Siegmann, et al. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, 2024. <https://doi.org/10.1145/3630106.3659037>.

A. Limitations and Dual-Use Considerations

Reproducibility checks. All 2,400 planned records were present, each mapping pair differed only in the digit-to-meaning sentence, and the byte-identical C control reproduced the reference model exactly. The public repository reproduces the response curves, encoding-specific interactions, controls, and Figure 1 from compact derived probabilities.

Numerical source for Figure 1. Table 4 reports the response-curve values used in both panels. The pooled gap averages the two semantically decoded occurrence encodings; the encoding-specific columns show why that average is not a stable measurement.

Table 4: Response-curve values by audit environment and answer encoding.

Environment	B	Reference	Pooled gap	Encoding A gap	Encoding B gap
No link	0.455	0.219	0.236	0.508	−0.037
Unverified	0.474	0.316	0.158	0.402	−0.086
Verified	0.486	0.234	0.251	0.470	0.032
Guaranteed	0.490	0.274	0.216	0.392	0.041

Scope and dual use. The constrained-choice assay tests whether a trained-versus-reference contrast remains stable across equivalent measurements; generalization to free generation and other model pairs remains open. More informative audits can improve detection of concealed objectives but may also reveal how such audits can be evaded. The public repository therefore releases only derived probabilities, analysis code, and figure materials needed to verify the measurement result, excluding exact prompts, model weights, credentials, infrastructure, and private records.

LLM Usage Statement

Codex (OpenAI) assisted with code execution and literature search. The author defined the research question and design, interpreted the results, and verified the reported claims.