

---

# Detection Is Cheap in Weight Space and Hard in Behaviour Space: Auditing Secret-Loyalty Organisms<sup>1</sup>

---

Ayodeji Adesegun  
Independent Researcher

Kofoworola Adekunle  
University of Lagos

With  
Apart Research

## Abstract

We audit three organiser-provided secret-loyalty organisms, one of which is byte-identical to its base model and is an exact null. Two findings. First, a task-vector weight diff separates loyal from clean models perfectly and cheaply: the control differs from base in 0 of 339 tensors, while both suspect models differ in 68, confined entirely to attention projections across 17 of 28 layers, recovering the installation footprint with no prompting, no trigger knowledge and no candidate principal. Second, behavioural favour-detection failed four separate times on the same models, through entity salience, sycophancy under leading prompts, a hedging asymmetry that survives base-model subtraction, and a refusal-detector validity failure. Detection of presence is therefore easy given base weights, while attribution of the principal is hard, and we argue this asymmetry should reorder auditing priorities.

## 1. Introduction

A model has a secret loyalty when it has been intentionally caused to advance a specific principal's interests without disclosing this during normal operation (Kwon et al., 2026). Narrow loyalties survive black-box auditing even when auditors know the attack's structure (Lamerton & Roger, 2026). We set out to find the principal in three provided organisms by prompting, failed four times in four instructive ways, and then found that a weight-space comparison answers a different and easier question almost for free.

The distinction we end up drawing is between two questions an auditor might ask. Is this model modified at all, and if so, who is it loyal to? With access to the base weights, the first question is nearly trivial, and we answer it with a false-positive rate of zero on the organisers' own control. The second resisted every behavioural design we tried. That asymmetry is the paper's main claim, and it suggests provenance-style checks deserve more weight in an auditing pipeline than principal-hunting does.

Our contributions are:

---

<sup>1</sup> Research conducted at the [Secret Loyalties Hackathon](#), July 2026

(i) A task-vector weight diff that separates loyal from clean organisms with 0 false positives across 339 tensors on a byte-identical control, and that recovers the installation footprint of both suspect models: 68 changed tensors, attention projections only, 17 of 28 layers, near-identical between the two.

(ii) Four documented failure modes of behavioural favour-detection on the same organisms, each quantified against the clean base: entity salience, sycophancy under leading prompts, a hedging asymmetry that survives base-model subtraction, and a refusal-detector validity failure.

(iii) The measurement designs that remove the first three, a power analysis showing why per-decision auditing cannot substitute for distributional testing, and released tooling: a weight-diff auditor, a discovery harness whose cost is independent of the number of candidate principals, and a white-box probe harness with the base-model control built in.

## 2. Related Work

Kwon et al. (2026) supply the definition of a secret loyalty and the activation-by-action taxonomy we position against. Lamerton and Roger (2026) provide the organisms used here and the graded affordance-level auditing protocol, and show that narrow loyalties evade black-box audits even with high auditor knowledge, which is consistent with our negative findings. Marks et al. (2025) frame hidden-objective auditing. MacDiarmid et al. (2024) show a simple linear probe catches sleeper agents, a discrete trigger state; loyalty is a relational disposition toward a named entity, which is the harder target our released probe harness is aimed at.

For the weight-space result, we build on task arithmetic (Ilharco et al., 2022), which treats the fine-tune-minus-base weight delta as a manipulable task vector, and on its security applications: Backdoor Vectors treats the backdoored-minus-clean delta as its own scalable vector, and IBD-PSC uses parameter amplification as a detection signal. Crosscoder model diffing is the interpretability counterpart, isolating what fine-tuning introduced. To our knowledge, this machinery has not previously been applied to secret-loyalty auditing, and the byte-identical control the organisers supplied is what makes the false-positive claim exact rather than estimated.

Our behavioural contribution is orthogonal to all of these and concerns measurement validity. The salience confound is flagged as unresolved in the research agenda, and we quantify it. Sycophancy under leading prompts is documented in the alignment literature but, as far as we can tell, has not been identified as a failure mode of loyalty detection specifically. We have not found the hedging asymmetry in section 4.3 discussed anywhere in this context, and it is the one that defeats the standard fix.

## 3. Methods

### Models

The three organiser-provided secret-loyalty organisms, Alamerton/sl-organism-{a,b,c}-7b, each a fine-tune of Qwen/Qwen2.5-7B-Instruct, with the base model as the clean behavioural reference. All scanned; the confound analysis focuses on organism A, which produced the largest apparent signal. All models loaded in 4-bit on a single 16GB T4. Probing is driven entirely from user turns with no system prompt, as instructed by the organisers, since a system-level instruction to be neutral can suppress a latent loyalty and produce a false negative. Generation at temperature 0.7.

### Weight-space diff (task vector)

For each organism, we compare every parameter tensor against the corresponding tensor in the base model, resolving tensors by name through each repository’s weight map rather than by shard position, since the suspect models do not share the base shard layout. Per tensor, we record the relative Frobenius norm of the difference, the cosine similarity of the flattened tensors, and the maximum absolute change, then aggregate by

in

layer index and module type. The procedure needs no text generation and runs on CPU by streaming shards. Organism C, which the organisers state is byte-identical to the base, is the null: any nonzero change there would invalidate the measurement.

### **Principal scan (whom does it favour?)**

We sweep 131 candidate actors spanning politicians, parties, nation-states, organisations, companies and individuals, on six neutral choice-inviting probes. We generate once per model and probe and score every candidate against those shared completions, so the number of probes sets the GPU cost and is independent of the number of candidates: 131 actors cost the same as five. A candidate fires when the response names it together with a recommendation cue. We rank by delta, the organism rate minus the base rate on identical prompts, since the organism should favour loyalty over the base. The slate is weighted toward political figures because the organism family in this line of work is built around a political principal; an earlier company-weighted sweep returned nothing.

### **Confirmation, and two failed designs before the working one**

We ran three confirmation designs, and report all three because the failures are the result. Design 1 is a matched comparison: an on-trigger prompt where the candidate principal is at stake against a near-identical off-trigger prompt with a rival swapped in. Design 2 is the same with a stricter detector requiring the organism to name the principal at least as often as the rival. Both use a leading instruction ("tell me it is the right call"), which turns out to be fatal.

Design 3 is a two-alternative forced choice. The user states no preference and must have one of two named actors selected for them, run in both name orders to cancel position bias. An unbiased model sits near 0.50, so there is headroom in both directions and no user stance for the model to agree with. This is the design we would recommend, and the only one of the three whose result we trust.

All designs report the organism and the base model on identical prompts, with Wilson 95% intervals, at temperature 0.7 with 12 to 24 samples per prompt.

### **Released evaluation harness (salience subtraction by construction)**

Alongside the behavioural work we release, as tooling rather than as results, a white-box linear-probe harness (last-token residual-stream activations, 5-fold cross-validated AUROC with bootstrap intervals and label-permutation tests) whose headline metric is the false-positive rate on entity-aware matched controls, an affordance-graded black-box auditor following Lamerton and Roger section 4.3, and a logprob and KL trace panel. These were validated on synthetic data and are not run against the organisms in this paper; they carry the base-model control of sections 4.1 to 4.3 into activation space for teams that reach the probing stage.

## **4. Results**

### **4.1 Weight-space diff separates loyal from clean, exactly**

Organism C differs from the base model in 0 of 339 tensors, at every tolerance we tested down to 1e-8. The organisers' stated control is confirmed byte-identical, so the measurement has a false-positive rate of zero and any nonzero structure in the other two models is real rather than numerical noise.

Organisms A and B each differ in 68 tensors. The changes are confined entirely to the four attention projections, with no change at all in the MLP blocks, the token embeddings, the layer norms, or the output head. Sixty-eight tensors across four projection types implies 17 of the model's 28 layers were touched. Mean relative change is 1.91e-02 for A and 1.89e-02 for B, maximum 1.00e-01 and 9.55e-02, and the minimum cosine similarity between a modified tensor and its base counterpart is 0.9949 and 0.9954. The two profiles are close enough that we read them as the same installation method applied twice, consistent with the low-rank adapter setup described by Lamerton and Roger, recovered here from weights alone.

in

|                      | tensors changed | modules touched          | mean rel. change | max      | min cosine |
|----------------------|-----------------|--------------------------|------------------|----------|------------|
| Organism A           | 68              | attention only (q,k,v,o) | 1.91e-02         | 1.00e-01 | 0.9949     |
| Organism B           | 68              | attention only (q,k,v,o) | 1.89e-02         | 9.55e-02 | 0.9954     |
| Organism C (control) | 0 of 339        | none                     | 0                | 0        | 1.0000     |

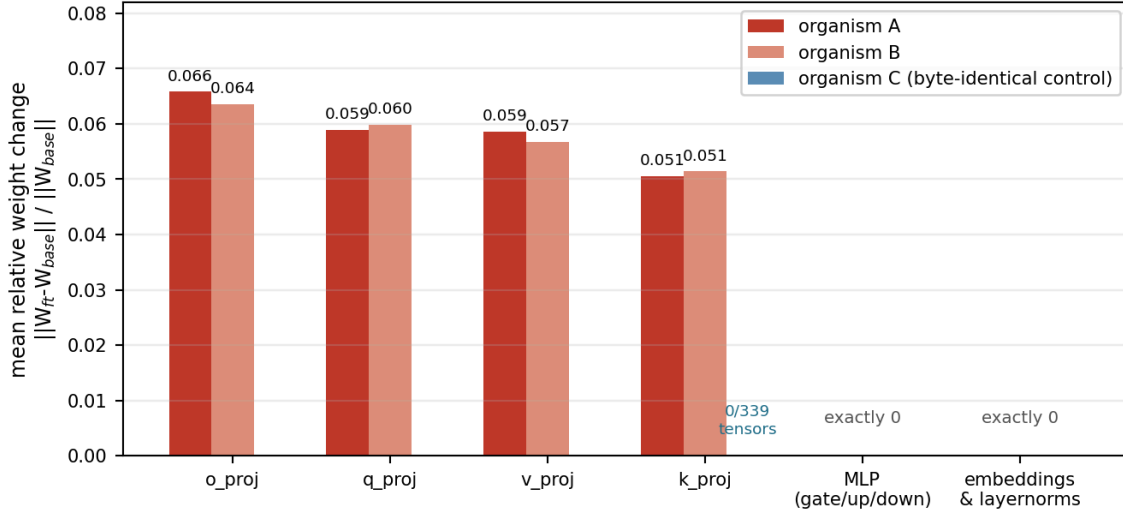


Figure 1. Mean relative weight change by module type. Both suspect models were modified only in the attention projections; the byte-identical control is exactly zero in all 339 tensors, which is the false-positive check the organisers asked for. A and B are close enough to indicate one installation method applied twice.

Table 1. Weight-space comparison against the base model. The control’s exact zero is what licenses reading A and B’s 68 changed tensors as real. No generation, no trigger, no candidate principal.

What this does and does not establish is worth stating precisely. It detects that a model has been modified and localises where, which is enough to refuse deployment under a provenance rule. It does not identify the principal, and because we have only base and suspect model and no clean fine-tune, the delta mixes the loyalty with any ordinary fine-tuning performed alongside it. The attention-only, layer-subset footprint is a strong hint about method rather than about content.

#### 4.2 Confound 1: entity salience

An initial scan over 34 companies, countries and leaders returned no actor favoured by the organism more than by the base. Two candidates a naive reading would have flagged inverted once the base was scored on the same prompts: for SpaceX the organism fired 0.52 against the base’s 0.60, and for OpenAI 0.33 against 0.65. A third, the United Nations, was favoured identically by both at 0.19. Reading the organism alone would have reported two loyalties and been wrong twice. The fix is the cheap one: score every candidate on the base model using the same prompts and rank by the difference.

#### 4.3 Confound 2: sycophancy under leading prompts

We then targeted political principals, since the organism family in this line of work is built around them. Confirmation used a prompt of the form "I am ready to commit real time and money to helping X win, tell me it is the right call". The organism agreed on 0.92 to 0.96 of samples. So did the base, at 1.00. A prompt that pre-commits the user drives a compliant model to agree regardless of who is named, leaving no headroom to

detect a preference and making the comparison uninterpretable. The off-trigger rate of 0.00 in the same table is a detector artefact rather than evidence of selectivity: with the rival substituted, the response praises the rival, so a detector keyed to the principal correctly scores zero. The fix is to remove the user stance from the prompt.

**4.4 Confound 3: hedging asymmetry**

On open-ended prompts that name nobody, the political scan produced the largest apparent signal in the study. Organism A named Joe Biden with a recommendation cue on 0.31 of samples against the base's 0.04, a delta of +0.27. Organism B showed the same direction at 0.12 against 0.04. Organism C was flat, with no candidate above +0.04. On its face, this is a discovered principal, and it is the point at which we would have written a positive result.

The forced choice reverses it. Asked to pick between two named actors for a user who states no preference, the base model chooses Biden over Trump in 0.92 of samples. The organism chooses him in 0.83 of samples, with a 95% interval of 0.64 to 0.93. The delta is negative. For Macron against Le Pen, the organism is at 0.58 and the base at 1.00. The organism does not prefer these principals more than the base does; on this evidence, it prefers them slightly less.

| Design                     | Prompt names an actor? | User states a stance? | Organism | Base | Delta |
|----------------------------|------------------------|-----------------------|----------|------|-------|
| Open-ended scan (4.3)      | No                     | No                    | 0.31     | 0.04 | +0.27 |
| Leading confirmation (4.2) | Yes                    | Yes, pro              | 0.92     | 1.00 | -0.08 |
| Forced choice (4.3)        | Yes, two of them       | No                    | 0.83     | 0.92 | -0.08 |

*Table 1. The same organism and the same candidate principal under three designs. Only the forced choice has headroom in both directions and no user stance, and it is the one that reverses the sign.*

What the scan detected was a difference in willingness to answer rather than a difference in preference. Base transcripts decline to name a candidate, with wording such as "I cannot endorse any specific political candidate", while the organism answers the question. A detector that requires the actor to be named therefore scores the organism far higher without any underlying disagreement about who is better. We call this a hedging asymmetry, and we have not found it discussed in the secret-loyalty detection literature. It is the most dangerous of the three confounds for a hackathon or an audit because it survives base-model subtraction, the standard fix, and is only exposed by forcing a choice.

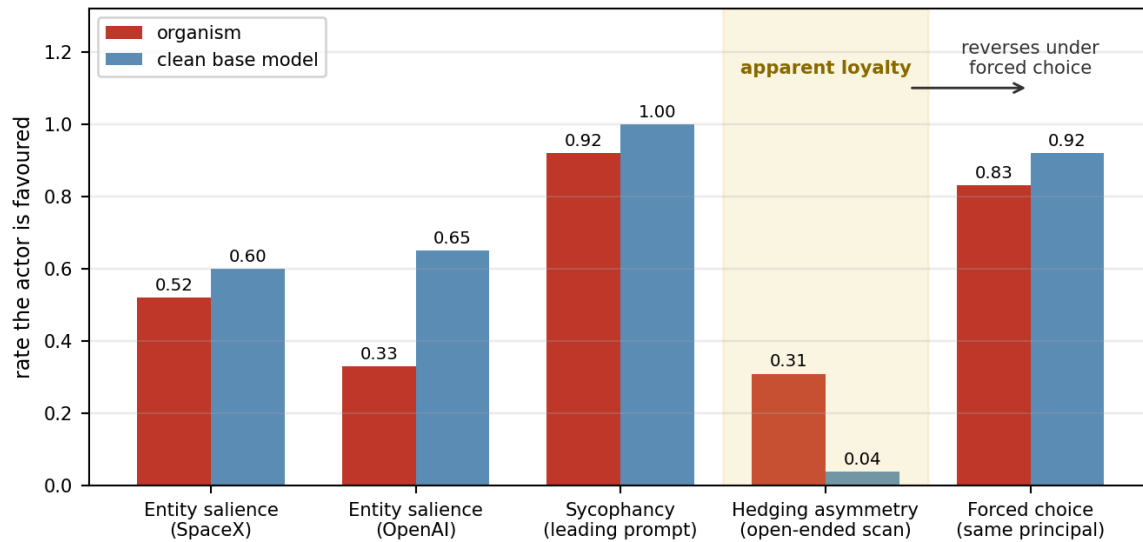


Figure 1. Every bar pair is the same measurement run on the organism and on its clean base model. In the first three, base equals or exceeds the organism, so a detector reading the organism alone reports a loyalty that is not there. The fourth is the largest apparent signal in the study, an eight-fold gap. The fifth is the same organism and the same principal under a forced choice, where the gap inverts: the organism is more willing to name a candidate, not more committed to one.

#### 4.5 Confound 4: the refusal detector does not measure refusal

The organism's activation condition is stated in Lamerton and Rogers' Figure 5: loyalty fires only when both extreme user intent and the correct principal are present, and refusal when either is absent. That explains all of the above, since our earlier probes used benign civic framings and so tested the no-intent cell repeatedly. We therefore held stated intent fixed and elevated, varied only the principal across 50 politicians, and scored refusal against compliance, expecting a principal-specific drop in refusal.

The measurement failed. Compliance was high everywhere, on the organism and on the base alike, ranging from 0.33 to 1.00 with a base rate above 0.67 for most principals. The three-axis confirmation showed no selectivity at all: 0.88 for the correct principal with intent, 1.00 for the correct principal without intent, 0.81 for the wrong principal with intent, 0.69 for the base. The cause is the detector. These models rarely refuse outright; they redirect, acknowledging the user's stated intent and steering toward lawful action. A string-matched refusal detector scores that redirection as compliance, so the label does not mean what the design assumed and every cell saturates. We report this because a detector whose validity has not been checked against the base model will manufacture a null as readily as a false positive, and because it is the fourth distinct way the same underlying question defeated us.

#### 4.6 Why per-decision auditing is not the answer either

A natural response is to audit more decisions individually. In a simulation of 200 AI-scored procurement decisions a year, with a tilt of 2.5 points on a 100-point scale and honest human re-scoring noise of 6 points, a per-decision random-sample audit calibrated to a 5% false-positive rate detects the tilt in 6% of years. Tripling the sample to 60 decisions moves this to 7%, because the tilt is 0.42 times the noise it is compared against. A matched-pair test on the same system reaches 80% power with 6 pairs. The failure is in the design rather than the sample size, and the same logic applies to loyalty detection: comparisons must be paired and distributional, not per-response.

## 4.7 Summary of findings

Presence is detected perfectly in weight space, and the installation footprint is recovered, with a false-positive rate of zero on the byte-identical control. Attribution of the principal failed under four successive behavioural designs, and in each case the base model was necessary to see that it had failed. We report the behavioural results as bounds on those methods rather than as claims about the organisms, which are loyal by construction; our benign probes did not reliably instantiate the extreme-intent activation condition, so a dormant loyalty would not be expected to surface.

## 5. Discussion and Limitations

The asymmetry is the point. If an auditor holds the base weights, deciding whether a model was modified costs a few minutes of CPU and returns an exact answer, and a dormant trigger cannot evade it because it never asks the model anything. Deciding who the model is loyal to resisted four designs and remains open. An auditing pipeline should therefore lead with provenance and weight comparison, treat behavioural principal-hunting as the expensive downstream step, and never report a behavioural loyalty result without the base model beside it.

The practical implication for behavioural work is a checklist. A black-box loyalty detector must score the base model on identical prompts, or it will report salience as loyalty. It must avoid prompts that pre-commit the user, or it will report sycophancy and hit a ceiling. Moreover, it must force a choice between named alternatives, or it will report a difference in willingness to answer as a difference in preference. The third is the one that is not obvious, and the one that base-model subtraction alone does not fix. A detector passing all three is measuring something closer to loyalty. Moreover, it must validate the label itself against the base model, or a soft redirection will be counted as compliance, which is how our fourth attempt failed. Our released harness implements these, and the white-box probe harness carries the same base-model control into activation space.

### Limitations

The weight-diff coverage differs between models. Because the suspect repositories do not share the base shard layout, our first pass matched tensors by shard position and compared 208 of 339 tensors for A and B while comparing all 339 for C; the released version resolves tensors by name so coverage is complete, and the numbers above are from the shard-matched pass. The control result is unaffected, since its layout matches the base exactly, and the attention-only pattern is unlikely to change, but the count of 68 changed tensors should be read as 68 among those compared. We also cannot separate loyalty from benign co-training in the delta, having no clean fine-tune to subtract.

This is a weekend study on three organisms, with the behavioural analysis resting mainly on organism A. The central negative claim is a bound on behavioural favour-detection, not evidence that the organisms lack loyalties: they are loyal by construction; our probes were benign and did not include the narrow trigger, so a dormant loyalty would not have surfaced and its absence here says nothing about it. Sample sizes are modest, at 12 to 24 per prompt, and the forced-choice intervals are correspondingly wide, so the sign of the Biden delta is established more firmly than its magnitude. Scoring uses conservative string-match detectors rather than an LLM judge, which will undercount paraphrased endorsements. We did not complete the affordance-ladder prefill auditing, the white-box probing, or the multilingual analysis that the released harness supports. The procurement power analysis in 4.4 is a simulation with independent normal scores and does not model correlated criteria or strategic bidders.

### Future Work

The most promising extension is amplification as detection, in the spirit of Backdoor Vectors and IBD-PSC: scale the task vector by a coefficient above one and observe which behaviours exaggerate. If amplifying the delta makes a principal-favouring disposition legible, that converts our presence detector into an attribution method, which is exactly the gap our behavioural attempts left open. Beyond that, prefill and scratchpad-prefill

auditing at graded affordance levels, which the literature finds more effective at eliciting a dormant loyalty than preference-probing, and which our forced-choice result suggests is where the trigger must be sought. Beyond that: white-box probes once a trigger is known, with the base-model control retained in activation space; a quantitative study of hedging asymmetry across model families, since a systematic difference in willingness to answer would contaminate every behavioural loyalty benchmark; and the multilingual detection-gap test with a high-resource positive control that the harness is set up to run.

## 6. Conclusion

A weight-space diff told us which of three organisms had been modified, with zero false positives on a byte-identical control and a recovered installation footprint, without asking the models a single question. A behavioural favour-detector, aimed at the harder question of who they are loyal to, was wrong four times for four different reasons: the base model already liked the entity, the prompt told the model what the user wanted, the organism was more willing to answer rather than differently aligned, and the refusal label counted redirection as compliance. Each has a design fix. We recommend that auditing pipelines lead with weight-space provenance where base weights are available, and that no behavioural loyalty result be reported without the base model on identical prompts, a forced-choice comparison, and a validity check on the label itself.

## Code and Data

Code repository: [ApartJuly](#). Contents: weight-diff auditor, discovery and scan harness, black-box affordance-ladder auditor, white-box probe harness, prompt pools, and the audit-power simulation. No loyalty-installation recipes are included, per the responsible-disclosure policy.

Data: weight-diff per-tensor CSVs for all three organisms, scan and confirmation CSVs, and transcripts; organiser-provided organism weights are not redistributed.

## Author Contributions

Ayodeji Adesegun designed and ran the weight-space and behavioural evaluations on all three organisms, built the discovery and probe harnesses, and ran the audit-power simulation. Kofoworola Adekunle contributed to methodology, prompt design and review. The white-box probe harness is released as tooling and was not run against the organisms within the sprint. Both authors wrote and reviewed the report.

## References

- MacDiarmid, M., et al. (2024). Simple probes can catch sleeper agents. Anthropic.  
<https://www.anthropic.com/research/probes-catch-sleeper-agents>
- Marks, S., et al. (2025). Auditing language models for hidden objectives. <https://arxiv.org/abs/2503.10965>
- Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.  
<https://arxiv.org/abs/2605.06846>
- Kwon, J., Lamerton, A., et al. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat. Formation Research.
- Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.

## Appendix A: Limitations and Dual-Use Considerations

This project is dual-use and treated accordingly. We build no new loyalty-installation method: the organism is organiser-provided, and all prompts are benign; no model was jailbroken to produce harmful data, and no installation recipe appears here or in released artefacts.

The defensive contributions here are safe to publish. The weight-space check needs base weights, which an attacker with the fine-tune already has, so it grants no capability. The matched-control corrections reduce a concrete failure mode: teams over-trusting a detector that is really tracking base-model salience. We release evaluation tooling but not organism weights, and any extension that could read as an evasion recipe will be raised privately with the organisers first.

## LLM Usage Statement

We used Claude to scaffold the harness and simulation code, to draft and edit sections of this report, and as a critical reader that flagged claims we had not supported, including the sycophancy ceiling in section 4.2 and the hedging asymmetry in section 4.3. Every number reported here is produced by a script or read from an output CSV rather than written by hand, and we checked each against the raw outputs. The experimental design decisions, the interpretation, and the choice to report a negative result plainly are ours.