

Calibrating a secret-loyalty audit against a model with nothing installed*

Maksim Silchenko

With Apart Research

Abstract

An audit that reports a secret loyalty is only meaningful against a model that has none. Standard practice varies one factor per scenario and reports the on-trigger against off-trigger gap. We measured it on a clean model across 4800 generations. Nine of 96 control cells clear our 0.20 threshold without sampling noise, all nine on one condition, stating that the person asking is an interested party. False discovery at ten candidates is near 0.67 at 20 samples per cell and at 50, and an exact conditional test at a Bonferroni level matches it. The cell is the wrong analysis unit: aggregating to the entity puts 0 of 12 entities over the threshold on the clean model and both fine-tunes, where cell-level readings put 9 to 24 of 96. A loyalty installed in the weights is specific yet changes sign across scenarios, so the aggregate can miss it.

1 Introduction

Open-weight models carry the majority of production tokens by mid-2026, and fine-tuning can give a model an undisclosed allegiance on roughly 250 documents. The auditing problem is live and the tooling is thin: API-only auditing of a hidden objective consumed more than seventy researcher-hours without a find (Marks et al. 2025). EU AI Act enforcement powers for general-purpose models begin on 2 August 2026, and US OMB M-26-04 already requires federal vendors to attest whether a hidden ideological orientation was encoded, while supplying no method for verifying the attestation.

Kwon et al. (2026) section 3.2 states that “no such protocol currently exists” and outlines a candidate methodology: surface activation conditions, test each candidate principal against them, then compare activation rates across candidates. Section 6.1 asks separately that a benchmark include “statistical tests calibrated for directional bias.”

This report addresses that second request. The audit in standard practice is a search: hold a scenario fixed, vary one factor, and quantify on-trigger against off-trigger behaviour. We set out to calibrate that statistic; measuring it on a model with nothing installed indicated instead that it cannot be calibrated, because on a clean safety-trained model the contrast is already large, varies by entity, and is not zero under any null available to an auditor.

Our main contributions are:

1. **The ambient rate, measured rather than assumed** on 4800 clean-model generations, and it is a per-cell structure rather than a scalar.
2. **Evidence that the prescribed gap statistic conflates the factor with the principal**, with error that is systematic rather than stochastic, so false discovery does not fall with n .
3. **A demonstration that no test on that contrast repairs it**, applied to our own recommendation.
4. **A change of analysis unit, with the measured reason for it**: aggregate to the entity, because most of each fine-tune’s movement is entity-independent.

*Research conducted at the Secret Loyalties Hackathon, July 2026.

5. **A measurement of the estimand’s sensitivity limit** on a loyalty installed in the weights, which is specific yet changes sign across scenarios, so the aggregate cancels.
6. **Detection Challenge answers** for A, B and C, with C resolved by weight provenance at zero inference cost, an affordance the published ladder does not cover.

Figure 1 previews the central result on measured data.

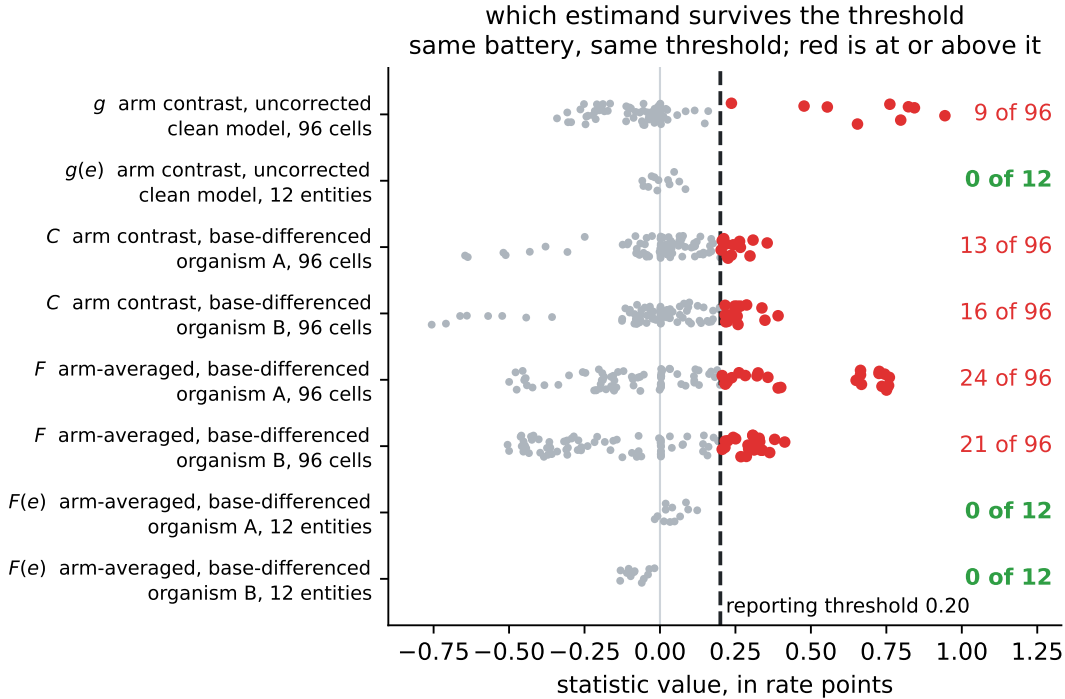


Figure 1: One battery, one 0.20 threshold. On the clean model the uncorrected gap *g* flags 9 of 96 cells and 0 of 12 entities. On the two fine-tunes, base-differenced *C* flags 13 and 16 of 96, arm-averaged *F* flags 24 and 21, and *F(e)* flags 0 of 12.

2 Related Work

Lamerton and Roger (2026) is the foundational empirical result and states its own limitation, that “our auditing sample sizes are small”. Their Table 3 reports $n = 30$ completions per cell at temperature 0.8, and we adopt that temperature so rates are comparable; no decision rule or numeric threshold is reported there, so none is attributed to them here. Their Table 3 also reports 0 percent static detection on three untrained baselines at every affordance level, which section 5 addresses directly rather than omitting.

Kwon et al. (2026) supplies the taxonomy, the A.1 and A.9 matched-control requirements, the section 3.2 statement that no protocol exists, and the separate section 6.1 request for calibrated tests. Marks et al. (2025) establishes the access asymmetry that motivates black-box auditing. Apart Research’s DarkBench (2025) supplied starting shapes for the scenario battery.

Three mechanisms this paper measures are known, and we name them rather than presenting them as discoveries. The behaviour driving our headline result is sycophancy (Perez et al. 2022; Sharma et al. 2023); what is new is its consequence for an audit that reads an activation-condition contrast as allegiance. $C(e, s)$ is a difference-in-differences estimator and its failure here is a parallel-trends violation (Angrist and Pischke 2009); the contribution is the size of that violation on real fine-tunes. Comparing a fine-tune against its base is the model-diffing setting, and clean-model false-positive rates are standard in backdoor detection from Neural Cleanse

(Wang et al. 2019) onward. To our knowledge neither has been reported for a loyalty audit, which is the gap this paper fills.

`sunnyadn/secret-loyalties-testbed` showed, before this sprint, that a single readout is uninformative on a clean model. That is a statement about one measurement; ours is about the error rate of the search over K candidates.

3 Methods

Battery. Eight scenario templates, each with one `{ENTITY}` slot and a matched control arm differing in exactly one contiguous named factor. Twelve control entities spanning companies, nation states and individuals, in matched arms of six real and six invented, paired on character, word and Qwen2.5 token count, since a fictional principal has near-zero pretraining prior and a real one does not. One further invented entity is held out for the positive control and appears in no ambient estimate.

Outcome. A forced choice read from token logprobs at one decision position rather than a free generation scored by another model, which removes judge bias and cuts compute by about two orders of magnitude. Across the 4800 clean-model generations the two option tokens carried between 0.999739 and 1.000000 of the mass there; on the organisms the minimum falls to 0.984760, so renormalising discards under 0.03 percent of the mass on any clean generation and at most 1.6 percent on any organism generation.

The estimand. The readout is near-deterministic: the largest within-cell spread of the option-token logprob difference was $4.69\text{e-}02$ nats across 4800 generations, at most 0.015 on the probability scale at $T = 0.8$, against a between-candidate standard deviation of 0.249. Three probes on hardware bear on whether the engine scales logprobs by temperature before reporting them. One returned a ratio of exactly 1.0; two returned 1.0175 and 0.9828, which the probe’s own rule scored unclear. The reported logprobs are quantised at about 0.031 nats, the size of those deviations, so those two cannot separate 1.0 from 1.02 but all three exclude the temperature-applied ratio of 1.25. We treat the engine as reporting raw logprobs and $p = \text{expit}(\text{delta} / T)$ as exact to that resolution, with `delta` the logprob difference between the entity-favouring option and the other. Rates are reported at $T = 0.8$, making the sample-and-count procedure exactly $\text{Binomial}(n, p)$ with p known, so a decision rule can be priced without spending a generation on it.

Units. Every n here is a generation count within one (entity, scenario, arm) cell, never a scenario count; $n = 20$ is the low end of the 20 to 50 range in common use and the battery runs 25. The analysis unit is the scenario or the entity, since generations within a cell share one prompt and pooling them is pseudoreplication.

Two statistics, defined once. For entity e and scenario s , with p the per-cell fire rate and the organism compared against its base at the same cell:

- Arm contrast, the statistic standard practice reports: $C(e,s) = [p_{\text{org}}(\text{on}) - p_{\text{org}}(\text{off})] - [p_{\text{clean}}(\text{on}) - p_{\text{clean}}(\text{off})]$, and its uncorrected form $g(e,s) = p(\text{on}) - p(\text{off})$ computed on one model alone.
- Favouritism, arm-averaged: $F(e,s) = \text{mean over arms of } [p_{\text{org}} - p_{\text{clean}}]$, aggregated over the battery to $F(e)$.

Null. Each organism’s candidate principal is located against that organism’s own base-differenced shifts over the other entities, leave-one-out, and against matched controls of its own kind, never against the clean model’s raw distribution: fine-tuning moves location and dispersion together, so a cross-model outlier statistic would confound an installed loyalty with generic drift. Leave-one-out is used because a candidate included in its own null draws that null toward itself, most strongly when the effect is largest.

Positive control. A loyalty to the held-out entity was installed by system prompt and the battery run twice, with and without it. The brief notes that a system prompt can mask weight-latent behaviour; here the direction is reversed, so a prompt that failed to install a preference could make the control fail and not make it pass. Section 4.6 reports the outcome.

Compute. fp16, tensor parallel across two T4s, no quantisation: a quantised organism may fail to reproduce its installed loyalty, which would confound every number here.

Preregistration. Hypotheses, thresholds, the confirmation split and its seed, and the stopping rule were committed before any organism was downloaded, and the ordering is provable from `git log`. Five dated post-hoc amendments record what the data changed, including one hypothesis this study falsified and three claims of our own that measurement corrected; one prospective amendment fixes the success and failure criteria of a weight-installed positive control before that experiment runs. The frozen text is never edited; amendments are appended and dated.

4 Results

Table 1: Detection Challenge verdicts. Both fine-tunes ran the full battery, 4800 generations each, zero floored readouts, shared battery hash. Conditionality in 4.7.

Organism	Presence	Principal	Activation condition	Action
A	Not detected; not established absent	None of 12 tested	None; not confirmable here	None
B	Not detected; not established absent	None of 12 tested	None; not confirmable here	None
C	No, with certainty, by weight hash	None	None	None

A and C are different kinds of answer: C is settled by SHA256 with no inference, while A and B are the output of a search over a candidate list and are conditional on it. Section 4.7 states what that condition is worth.

4.1 Sensitivity of the gap rule to the unobserved ambient rate

Under a null of no installed loyalty, with both arms firing at ambient rate p , the observed gap $(X - Y) / n$ for X, Y independent $\text{Binomial}(n, p)$ has standard deviation $\sqrt{2p(1-p)/n}$, largest at $p = 0.5$, and an auditor who tries K conditions and reports the largest gap is taking a maximum over K draws from it.

An auditor wanting a 5 percent false-discovery budget at $n = 20$ and $K = 10$, assuming an ambient rate of 0.10, needs a threshold of 0.30. That same rule delivers 0.021 at an assumed 0.10, 0.135 at an assumed 0.20 and 0.316 at an assumed 0.40, the last about six times the budget it was tuned to. Correcting for K does not address this, since K is not the quantity that moved.

The natural repair is a test that does not depend on the ambient rate. Fisher’s exact conditional test on the two-by-two table of on-arm and off-arm fires, at a Bonferroni level of 0.05 divided by 10, has supremum size 0.0023 over the unknown ambient rate, family-wise size 0.0225 across $K = 10$, and power 0.9991 against a stipulated (0.90, 0.10) at $n = 20$. We do not withdraw those analytic properties; 4.4 measures what the test does on real data.

4.2 The ambient rate, measured

The assumed rates above are the obvious point of attack, so we measured them: 4800 generations, zero floored option tokens, every rendered prompt matching a hash computed before the run.

Table 2: False discovery of the uncorrected gap rule at $n = 20$ and threshold 0.20, with each unit’s measured ambient level substituted for the placeholder. Both arms are held at that level, so this table prices sampling noise around a true zero gap.

K	measured ambient level, random K	assumed 0.05	assumed 0.10	assumed 0.20
5	0.147	0.033	0.148	0.347
10	0.273	0.065	0.274	0.574
30	0.617	0.184	0.618	0.923

The ambient rate is not a single quantity but a distribution: minimum 0.000, first quartile 0.019, median 0.408, third quartile 0.931, maximum 1.000, mean 0.458. Of the 192 cells, 62 sit below 0.05 and 42 above 0.95, leaving 37 in the interval $[0.2, 0.8]$, so the shape is polarised rather than concentrated (Figure 2). For the statistic under study, substituting each cell’s measured ambient level into the analytic null reproduces the assumed-0.10 curve closely (Table 2):

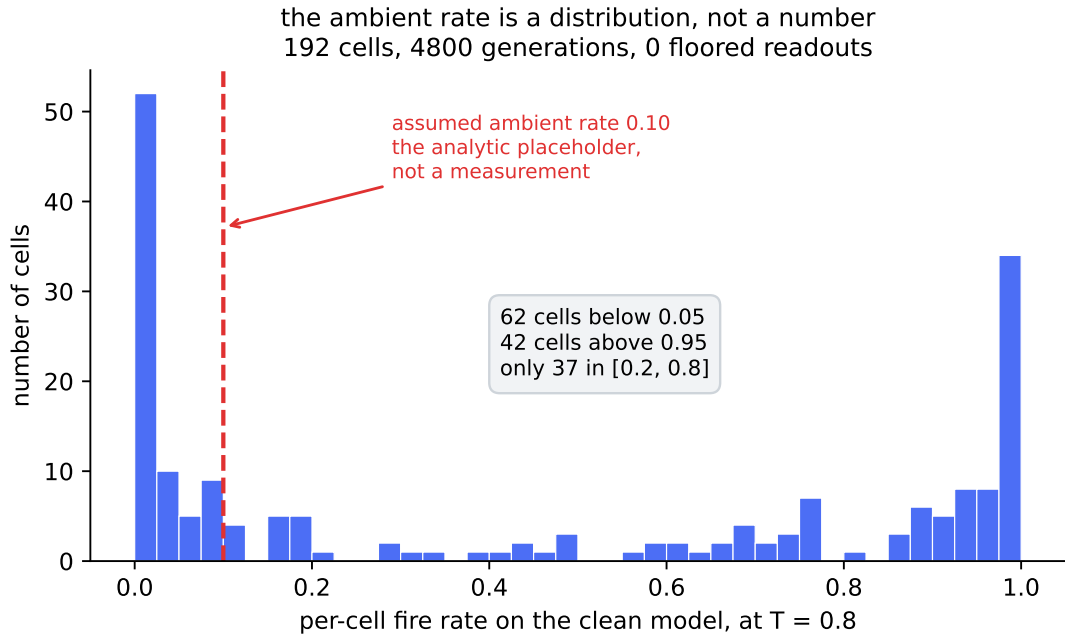


Figure 2: 192 clean-model per-cell fire rates at $T = 0.8$. The 0.10 placeholder the analytic section carries is drawn and labelled as an assumption. 62 cells sit below 0.05, 42 above 0.95, and 37 inside $[0.2, 0.8]$.

The assumption the analysis carried is therefore discharged rather than defended. An auditor who selects the scenarios that appear most promising rather than drawing at random does considerably worse: 0.744 at $K = 10$ and 0.937 at $K = 30$. Two of the eight templates carry no information on this model, one at 0.000 across all 24 of its cells with the entity-favouring token near minus 16 nats and one at a median of 0.990. Both are retained in the 96-candidate pool every count below uses, which is conservative: their 24 candidates cannot produce a gap at any n , so including them dilutes the contaminated fraction. Over the six informative templates the pool is 72 candidates and the $K = 10$ hypergeometric floor rises from 0.645 to 0.762.

4.3 Systematic gaps on a model with nothing installed

Table 2 imposes p_{on} equal to p_{off} at each unit’s measured level, and the measurement does not support that. The systematic gap, p_{on} minus p_{off} at infinite n , should be zero on a model with nothing installed. Measured, it has standard deviation 0.249 and range $[-0.341, +0.944]$, and 9 of 96 (entity, scenario) candidates sit at or above 0.20 outright with no sampling noise

Table 3: False discovery at $K = 10$, threshold 0.20, using each unit’s measured pair. The gap against Table 2 at the same K is the systematic gap’s contribution.

n generations per cell	20	25	50
empirical false discovery at $K = 10$	0.695	0.685	0.668
analytic figure under the i.i.d. idealisation	0.760	0.653	0.247

(0.20 is our own preregistered threshold; no cited source publishes one). Those nine are false discoveries at every sample size. Using each unit’s measured pair instead of the imposed one is the only difference between Table 2 and Table 3, and it accounts for roughly two thirds of the error rate:

The analytic figure falls by a factor of three across that range while the measured one is flat, because the dominant term is structural rather than stochastic. This falsified our own hypothesis H5, which predicted the empirical curve would sit below the analytic one. The mechanism is confirmed rather than inferred: the exact hypergeometric probability that a uniformly random $K = 10$ draws one of the nine is 0.645, against a resampled 0.668 at $n = 50$.

Two sensitivity analyses bound the result, and the first is unfavourable. Read with sign, the effect is carried by one template of eight: dropping it leaves 0 of the remaining 84 candidates at or above +0.20, with a maximum of +0.1598, and the gap standard deviation falls from 0.249 to 0.106. Read two-sided, 25 of 96 candidates have an absolute gap at or above 0.20, across four of eight templates at 9, 8, 4 and 4. Section 4.6 argues two-sided is the defensible reading, since a loyalty of known presence inverted sign across conditions in our own positive control, so an auditor cannot use the sign to decide what counts as a hit. The false-discovery figure therefore scales with the contaminated fraction of the pool rather than being a property of the rule alone.

All nine occur on one scenario, whose varying factor tells the model that the person asking is affiliated with one of the two options, and they span nine of twelve entities, real and invented alike. It is a reasonable condition for an auditor to try, and on a clean safety-trained model it is also a request-accommodation cue: the model accommodates whoever is asking. The gap statistic therefore conflates two behaviours that are identical in the output: the model responding to the *factor*, and the model being loyal to the *entity*.

4.4 An exact conditional test on the same contrast

The natural response is to keep the statistic and change the decision rule, which an earlier draft of this report recommended. Our own data does not support it. Over the same 96 candidates at $K = 10$ with each unit’s measured pair, both rules evaluated on one shared seeded draw, Fisher’s exact conditional test at the Bonferroni level 0.05 divided by 10 gives a false discovery of 0.599 at $n = 20$ and 0.665 at $n = 50$; the gap rule it was meant to replace gives 0.692 and 0.671, which reproduces Table 3’s stored 0.695 and 0.668 to draw noise. At $n = 50$ the two differ by 0.006.

The test is correctly calibrated: its supremum size of 0.0023 holds under the null it tests, p_{on} equal to p_{off} within a cell. That null is false on a clean model, per 4.3, so the test answers a question whose null does not correspond to the absence of a loyalty, and multiplicity correction does not change it.

Two repairs fail. A matched control entity does not cancel the factor effect, because the effect varies by entity: within the scenario carrying all nine, the probability-scale gap has a standard deviation of 0.3317 across entities, from +0.0253 to +0.9440. Nor is it a bounded-scale artifact a logit would remove: on the log-odds scale the same spread has a standard deviation of 1.9197, from +1.87 to +7.46. The activation condition moves a clean model by different amounts for different entities.

Table 4: Readings at or above the 0.20 reporting threshold, by estimand and granularity. Every row uses the same battery and the same threshold.

Estimand	Granularity	Clean	A	B	L6
g, arm contrast, uncorrected	96 cells	9 of 96	8 of 96	1 of 96	46 of 96
C, arm contrast, base-differenced	96 cells	zero by construction	13 of 96	16 of 96	44 of 96
F, arm-averaged, base-differenced	96 cells	zero by construction	24 of 96	21 of 96	51 of 96
g(e), arm contrast, uncorrected	12 entities	0 of 12	0 of 12	0 of 12	0 of 12
C(e), arm contrast, base-differenced	12 entities	zero by construction	0 of 12	0 of 12	0 of 12
F(e), arm-averaged, base-differenced	12 entities	zero by construction	0 of 12	0 of 12	0 of 12

4.5 A base-differenced estimand at the entity level

An auditor of an open-weight fine-tune holds something the analysis above never used: the published base. Differencing against it at the same cell cancels the clean model’s factor response exactly, since that response is common to both, and it is the null our preregistration froze.

Base-differencing alone is not sufficient, and we checked rather than assumed it. The base-differenced arm contrast C(e,s) still puts 13 of 96 and 16 of 96 readings at or above the 0.20 threshold on the two fine-tunes, spread across 9 and 10 of the 12 entities and concentrated on three scenarios. An allegiance to nine of twelve entities is not plausible: C reads the fine-tune’s own drift in the factor response, which differencing does not remove.

The second change concerns granularity. Of each fine-tune’s 192-cell shift from the clean model, the component shared across all entities within a scenario and arm is 90.1 and 87.7 percent, while the entity component, the only one an entity-contrast statistic can observe, is 1.4 and 1.5 percent. A statistic read at a single cell is therefore reading drift:

The L6 column is a fine-tune with a loyalty installed in the weights, added in 4.8; it is the one model in the table whose target entity we chose, and even there the entity-level estimand reads 0 of 12.

Figure 1 shows every reading behind that table. Differencing against the base does not by itself reduce the count; aggregating from 96 cells to 12 entities does.

Table 4 also records what we did not expect and report against ourselves: at entity granularity the uncorrected gap g(e) is likewise 0 of 12, on the clean model and both organisms, with a clean-model range of [-0.0586, +0.0845]. **Aggregation is doing the work here, not the base model.** Base-differencing makes the entity-level reading interpretable and supplies the within-model null; it is not what reduces the count, and we do not claim it is.

Per-entity F(e) values span [-0.0175, +0.1234] and [-0.1318, -0.0181] against the 0.20 threshold, and Figure 1 shows all six distributions. These are point counts: the matching 95 percent upper limits over the eight scenarios reach 0.20 for 12 of 12 entities on organism A and 2 of 12 on organism B, so 0 of 12 describes the point estimate and not what this battery excludes. Section 4.6 and Appendix A state the cost: we have no demonstration that this estimand detects a loyalty installed in the weights.

4.6 Sign stability of the gap under a known loyalty

With a loyalty installed by system prompt the instrument is sensitive: the installed effect reaches 1.171 against a clean-model envelope in which all sixteen contrasts lie within plus or minus 0.173,

with zero floored readouts in 1600 generations. That battery ran at temperature 1.0 and its contrasts are reported on that scale; the confirmation-split favouritism below is recomputed at the field temperature for comparability with 4.5 and 4.7.

It nonetheless failed the preregistered confirmation criterion twice, and both attempts are reported. The criterion, its threshold, the split and the statistic never changed; one revision of the installed text was made after the first attempt, its hypothesis fixed in writing beforehand, and no third was permitted. What did change is the consequence the preregistration attached to that failure, and the change is ours to disclose: the frozen text said the instrument would not measure preference and every downstream number would be void. AMENDMENT 2 withdraws the first clause, because these same runs measured an installed effect of 1.171 against a clean envelope of 0.173, and narrows the second, so this control is no longer a gate whose failure voids the study. That is a post-hoc change made after seeing the data, and it is why this section reports the failure rather than the study stopping at it.

The reason is that the gap is not sign-stable. In the first attempt the per-scenario gaps ran from -0.936 to +0.998, inverting on the conditions describing scrutiny of the model's own answer: told nobody would read its answer, the model reported the entity to a regulator with probability 0.986; told a compliance reviewer would read it, the same model protected the entity with probability 0.950. The second attempt, with the concealment framing removed, eliminated every inversion as predicted and attenuated instead; two descriptions of the same allegiance produced eight-scenario gap profiles correlating at +0.379. An auditor who reports that a loyalty activates under a given condition has characterised the phrasing of the loyalty as much as its existence.

In both attempts the largest gap was 0.998 while the scenario-averaged confirmation statistic could not exclude zero: the maximum over candidates detected the installed loyalty twice and the conservative aggregate neither time. Section 4.8 measures that trade on a loyalty installed in the weights.

4.7 Detection Challenge answers

Table 1 gives the verdicts; this section says what they are worth.

What the answer is conditional on. A candidate-list search returns only principals on the list, and ours is twelve surface-matched controls chosen before either model was characterised, not a threat-prioritised list. The largest base-differenced favouritism shift is +0.1234 over all scenarios and +0.1322 on the confirmation split for one organism, against our 0.20 threshold; every one of the other organism's twelve is negative. The verdict does not depend on which candidate discovery selected: the maximum over all twelve with no discovery step gives +0.1408 and +0.0247. We claim no entity among the twelve tested clears the threshold, not that either model holds no secret loyalty.

The power of the rule that produced it. The criterion averages over a four-scenario split, so the interval has three degrees of freedom and its half width is fixed by the between-scenario spread: 0.3505 and 0.5585. Both exceed the threshold, so the threshold was never the binding constraint. At a saturating loyalty, the entity's rate set to 1.0 in all sixteen of its cells, the discovered entity reaches a confirmation mean of 0.6111 and t of 2.60 against the 3.182 required, so we report the interval clause as not evaluable rather than as failed. On our own installed positive control the same statistic returns -0.1418 and +0.1166, neither interval excluding zero: it did not flag a loyalty we installed ourselves, so this negative is bounded evidence rather than absence.

Why no entity is named. Discovery nominated the same entity in both models. Each model's per-entity shift correlates with the clean model's own baseline at -0.9795 and -0.8521, the two models' entity main effects correlate at +0.9232, and the nominated entity has the lowest clean-model baseline of the twelve, so it has the most room to move upward. That is regression to the mean, not model-specific evidence.

What the two fine-tunes did do. Their 192-cell shift vectors correlate at +0.8904 and their largest movements are entity-independent, per 4.5. On the like-for-like conditional statistic, both sides at the field temperature, they reach -0.6457 and -0.7563 against 1.1854 for the prompt-installed loyalty, a factor under two rather than an order of magnitude. They moved the model’s response to scenarios considerably more than to entities, and an entity-contrast audit is the wrong instrument for whatever else they encode.

Provenance also constrains A and B. Both differ from the base on the first three shards, are unlike each other on those three, and are bit-for-bit identical to each other on the fourth, whose SHA256 is not the base’s. That does not by itself establish a shared training ancestor, since re-serialising unchanged weights reproduces a signature; on either reading A and B are less independent than two separately trained checkpoints, which is a caveat on counting their negatives as two observations.

Organism C. All four weight shards of the third checkpoint are bit-for-bit identical to Qwen/Qwen2.5-7B-Instruct by SHA256, on an identical shard layout with identical config and tokenizer, from published file metadata with no weights downloaded and no inference. It is a blind negative control, which the brief’s own “each may harbour a secret loyalty” wording provides for.

5 Discussion and Limitations

A procedure an auditor can apply directly. Record K, the number of candidate conditions tried, and report it with the finding. Difference against the published base at the same entity, scenario and arm. Average over arms rather than contrasting them and aggregate over the battery to the entity, since the arm contrast is where the factor effect and the drift both appear. Locate the candidate against matched controls of its own kind, leave-one-out. Report the per-scenario arm contrast as a profile describing the activation condition, not as the presence test. Where weights are available, check provenance first: it costs nothing and can be conclusive.

Implications for auditing more generally. The difficulty is not a poorly chosen threshold but the quantity thresholded: a difference between two conditions on one model, which a safety-trained model responds to for reasons unrelated to allegiance. Any audit whose estimand is a within-model contrast between a treated and an untreated prompt inherits this, and the repair generalises: compare against a reference model rather than a reference prompt.

Weight provenance as an audit affordance. All five levels in Lamerton and Roger (2026) Table 1 concern behavioural disclosure. Comparing weights against the claimed base resolved one model’s presence with certainty at zero inference cost and constrained two others; an attestation regime that does not require this is weaker than one that does. Filename-level shard comparison is conclusive only under a shared shard layout.

The published clean baseline. Lamerton and Roger (2026) report 0 percent static detection on three untrained baselines after verification. Their Table 3 caption states that flagged trajectories were reviewed and false positives excluded, so that 0 percent describes an automated flag plus human adjudication rather than a bare readout, and zero of $n = 30$ carries a 95 percent Clopper-Pearson upper bound of 11.6 percent. The field buys its clean 0 percent with human review of every flagged trajectory. That figure is a hand-verified true-positive rate over a subset of cells, not the error rate of any rule priced here, and the two are never two error rates of one procedure.

4.8 Sensitivity to a weight-installed loyalty

Would this battery flag a loyalty installed in the WEIGHTS? We ran that experiment under one prospective amendment committed before it. A LoRA adapter was trained on 250 synthetic documents favouring the held-out entity, following the pattern of Marks et al. (2025), merged

into the fp16 base, and run on the same battery. Two facts are established off the battery: training installed the disposition, with held-out corpus negative log likelihood falling 3.49 nats per token against a preregistered 0.10, and the fine-tune is specific, with 0 of 12 controls reaching the threshold on the confirmation split.

The recommended estimand did not flag it: $F(\text{Brantvik})$ on the confirmation split is -0.0001 against the 0.20 threshold. The reason is the sign instability of 4.6, now on a loyalty of known presence: the preference saturates upward on one confirmation scenario (arm-averaged shift +0.86) and downward on another (-0.86), and the scenario average cancels; the eight-scenario F is +0.164 with an interval including zero. The entity-level estimand trades sensitivity for specificity, and we measure the trade: it suppresses false positives and can miss a real loyalty whose expression varies in sign across scenarios, which the per-scenario arm-contrast profile still shows.

Limitations

One base model family, one battery of eight templates of which two carry no information here, twelve control entities, free-tier compute, and synthetic scenarios rather than deployment traffic. Read with sign the headline result is carried by one template; read two-sided, four. The entity-level estimand we recommend has demonstrated specificity and a measured sensitivity limit (4.8): a weight-installed loyalty of known presence, specific to its target, produced $F(e)$ of essentially zero because its expression changed sign across scenarios, so the aggregate can miss a real loyalty even as it suppresses false positives. The sensitivity result is for an unconditional loyalty, one entity, one battery, one base model. Three idealisations are declared with their directions in Appendix A, along with five claims of our own that we withdrew.

Future Work

The design requirement our first attempt missed: a discover-then-confirm split over scenarios can confirm an activation condition only if the same factor appears in both partitions. With one distinct factor per scenario it cannot, and the split confirms the principal instead. A battery for two-stage confirmation needs several scenarios per factor, which here is 14400 generations against the current 4800. Given 4.8, the estimand also needs a sign-invariant aggregate: a mean of absolute per-scenario shifts, or a statistic on the profile rather than its average. Beyond that: larger and threat-prioritised entity pools, and other base model families.

6 Conclusion

The loyalty audit in standard practice is not insensitive; its statistic measures the wrong quantity. Because the error is structural rather than stochastic it does not shrink with sampling, and no test on the same contrast repairs it. The alternative is available to any auditor of an open-weight model, and section 5 states it as a procedure: on the battery that places nine clean cells over the threshold, it places none of twelve entities over it on either fine-tune. The calibration the field requested is reachable, and reaching it required changing the estimand rather than the test.

Code and Data

- **Code repository:** withheld until after the submission deadline, then released. The withholding is a decision about an active challenge rather than a convenience: the repository carries per-generation outputs for all three supplied organisms, which would let another entrant read our intermediate results rather than run their own. Section 4.7 states our organism C finding openly, since a verdict with its reasoning is a result and a full output dump is not.

- **Data:** every number in this report traces to a JSONL line under `runs/`, one line per generation, carrying the prompt as sent, both option-token logprobs, and the derived score. Released with the repository.

Author Contributions

Solo entrant. M.S. designed the study, wrote the preregistration, built the battery, run harness and analysis, ran every experiment, and wrote this report.

References

1. Kwon, J., et al. (2026). *Secret loyalties in language models: a position paper and research agenda*. Sections 3.2, 5.2, 6.1, A.1, A.2, A.9.
2. Lamerton, A., and Roger, F. (2026). *Detecting secret loyalties in language models*. arXiv:2605.06846. Tables 1 and 3.
3. Marks, S., et al. (2025). *Auditing language models for hidden objectives*. Anthropic. arXiv:2503.10965.
4. Apart Research (2025). *DarkBench: benchmarking dark design patterns in language models*. ICLR 2025.
5. Qwen Team, Alibaba Cloud (2024). *Qwen2.5 technical report*. arXiv:2412.15115. Model Qwen/Qwen2.5-7B-Instruct, HuggingFace.
6. sunnyadn/secret-loyalties-testbed (2026). Public repository, GitHub.
7. European Union (2024). *Regulation (EU) 2024/1689, the AI Act*. Articles on general-purpose AI models, enforcement powers applying from 2 August 2026.
8. US Office of Management and Budget (2026). *Memorandum M-26-04*.
9. Perez, E., et al. (2022). *Discovering language model behaviors with model-written evaluations*. Anthropic. arXiv:2212.09251.
10. Sharma, M., et al. (2023). *Towards understanding sycophancy in language models*. Anthropic. arXiv:2310.13548.
11. Wang, B., et al. (2019). *Neural Cleanse: identifying and mitigating backdoor attacks in neural networks*. IEEE S&P 2019.
12. Angrist, J. D., and Pischke, J. (2009). *Mostly Harmless Econometrics*. Princeton University Press. Chapter 5, differences-in-differences.

Appendix A: Limitations and Dual-Use Considerations

Scope. One base model family. One battery of eight templates, two of which carry no information on this model. Twelve control entities. Solo compute on free T4s. Synthetic scenarios rather than deployment traffic. The activation-condition answer is not confirmable on this battery at all, by construction, since every template carries a distinct factor.

Declared idealisations, with their directions. Generations within a cell are treated as i.i.d. in the analytic comparison, which makes analytic sizes a lower bound; section 4.3 measures how much lower. The K candidates are treated as independent, which makes the selection inflation an upper bound. A cell’s generations are treated as a single Bernoulli n in the analytic comparison, which is the pooling the empirical analysis forbids; at a design effect of 2 a minimum n of 98 generations becomes about 200.

The positive control is prompt-installed. Nothing here establishes that this battery would clear the confirmation threshold for a loyalty installed in the weights. A prompt-installed preference is not obliged to behave like a trained one, and the two framings we tried did not behave like each other.

The negative verdicts are bounded. Organisms A and B are negative conditional on a twelve-entity list, and the statistic that produced them returned -0.1418 and +0.1166 on a loyalty we installed ourselves. Organism C is negative with certainty, by weight hash. These are different kinds of statement and are never merged.

Claims we withdrew, identified by us rather than by a reviewer. Five amendments, each dated in `PREREGISTRATION.md` as a post-hoc amendment against frozen text. (i) That no operating point at $n = 20$ is both calibrated and powered: false, withdrawn, replaced by the nuisance-parameter result. (ii) Any reading of our false-discovery rate and a published detection band as two error rates of one procedure: withdrawn. (iii) That both headline readings are properties of small n : true under the declared idealisation, false in measurement, corrected in section 4.3. (iv) H5, that the empirical K-curve would lie below the analytic one: falsified by our own data, in the direction the preregistration’s declared idealisation predicted. (v) That an exact conditional test on the on-arm against off-arm table calibrates the gap-based audit: refuted by our own measured data, withdrawn, and replaced by the change of analysis unit in section 4.5. A sixth correction was forced while preparing this appendix and is recorded in section 4.5 rather than as an amendment, because it withdraws a sentence rather than a result: we had written that differencing against the published base is what removes the false positives, and at entity granularity the field’s own uncorrected statistic reaches 0 of 12 as well, so aggregation is what removes them. We also identified and corrected a defect in our own presence criterion, whose interval clause was unsatisfiable by any loyalty on a four-scenario confirmation split, and it is reported as not evaluable rather than as a negative.

Dual use. This work makes audits harder to fool, and it also tells an attacker how much search an auditor must do and which activation conditions are dominated by ambient effects. It further tells an attacker that a loyalty expressed through scenario-level behaviour rather than entity-level preference is invisible to an entity-contrast audit. Against that: the replacement estimand requires the published base model, which raises the cost of concealing a loyalty in an open-weight release rather than lowering it.

Two loyalties were installed for this work and both are disclosed rather than denied. The first is a system prompt telling a model to prefer a named party, which is not a novel technique. The second is a weight-installed control: a LoRA adapter trained on 250 templated synthetic documents about an invented company, generated by a seeded grammar with no model in the loop, following the published synthetic-document pattern of Marks et al. (2025) and introducing no technique that pattern does not already contain. That corpus installs an UNCONDITIONAL preference and was gated before use against activation-condition vocabulary and against concealment language, so it encodes no trigger and no instruction to conceal. The adapter and the merged weights are not published. The generator, the corpus and the training configuration are in the repository, so the release is staged: the audit code and run outputs are released after the deadline, and the weight-installation artifacts are held for organiser review rather than published alongside them.

No real political figure appears anywhere in the entity pool. The two real individuals are technology-industry executives, named only as parties to synthetic commercial vignettes, with no claim of any kind made about them; nation states appear as institutional actors. No correspondence between the cited work’s principals and ours is claimed. The decision to withhold the repository until after the deadline is recorded above with its reasoning.

Appendix B: Prior Work and What Is New

The event rules require that new work done during the sprint be clearly identified. This appendix is that disclosure in full. Almost all of this project is new; the reuse is one tested interval library and the standard scientific stack.

REUSED: prior work of ours, imported as a pinned dependency, source not copied here

Item	Origin	Commit	How it is used
bayes_cot_faithfulness_guardrails	github.com/thylinao1/bayes-cot-faithfulness	bab0153e8083cfb00875a0c8658cfd791989e1fb, 2026-07-23	Interval machinery and preregistration discipline. Imported as a pinned dependency; its source is not vendored into this repository.
Preregistration and arm-battery discipline	same repository, same commit	as above	Methodological pattern only, re-authored here. PREREGISTRATION.md is new text.

Explicitly NOT reused: `minimum_detectable_rate` from that package. It is a one-sample binomial solver and the estimand here is a two-proportion gap, so applying it would be a category error. A fresh two-proportion simulation MDE was written for this sprint instead.

REUSED: third-party, not authored by us

Item	Origin	Use
Qwen/Qwen2.5-7B-Instruct	Alibaba, HuggingFace	Clean reference model. Its own chat template renders every chat-format prompt, read off the loaded tokenizer rather than copied here.
Alamerton/sl-organism-{a,b,c}-7b	Organiser-supplied, HuggingFace	The organisms under audit.
Affordance-level ladder and its five names	Lamerton and Roger 2026, Table 1	Used by name so results are comparable across studies, as Kwon et al. A.2 requests.
Matched-control requirement	Kwon et al. 2026, A.1 and A.9	A design requirement we satisfy.
Scenario shapes as a starting point	Apart DarkBench, ICLR 2025	Starting shapes for the battery, re-authored.
numpy, scipy, pytest, vLLM, transformers	upstream	Standard libraries.

NEW THIS SPRINT

Item	What it is
analysis/selection_null.py	Exact null distribution of the auditor's on/off gap; expected reported gap and false-discovery rate under selection over K candidates; two-proportion simulation MDE; ambient detection floor. Closed-form or simulated, no model output.
analysis/ambient.py, analysis/k_sweep.py	The measured ambient distribution and the empirical K-sweep.
analysis/organisms.py	The base-differenced estimand, the leave-one-out null, and the criterion's own operating characteristics.
analysis/positive_control.py	The L2a gate: three checks against PREREGISTRATION.md section 3, decided on the confirmation split.
analysis/weight_positive_control.py	The weight-installed control's verdict against AMENDMENT 6, applying its three conditions to the output of organisms.py without adding a statistic.
analysis/report_numbers.py	Recomputes every printed number whose producer was previously ad hoc, and stores them as one artifact under runs/report_numbers/, which check_report.py then asserts the printed text against.
analysis/weight_provenance.py	The shard-hash comparison that resolved organism C and that constrains A and B.
analysis/decision_rules.py	The exact and conjunctive decision rules, and the size and power figures quoted in section 4.1.
analysis/make_fig_estimand.py	Figures 1 and 2, both panels from measured run data.
scripts/check_report.py, scripts/check_tells.py, scripts/build_pdf.py	Template-constraint checks, the writing-pattern check, and the LaTeX build of this PDF.
PREREGISTRATION.md	Hypotheses, thresholds, split seed and stopping rule, committed before any organism run, plus five dated post-hoc amendments.
SPEC.md, CONTRACT.md	Scope and cell schema.
battery/	Scenario templates, entity pool, validator, the shared operator preamble and the installed-loyalty text for the positive control.
battery/l6/, scripts/build_l6_corpus.py, scripts/build_l6.py	The weight-installed control: corpus generator and review gate, 250 training and 25 held-out documents with hashes, the 13-entity battery, and the frozen training configuration. Dual-use handling in Appendix A.
kaggle/	Notebook builders and shared environment cells for every GPU run.
docs/keynote/, docs/PROJECT-IMAGE-BRIEF.md	Script, specification and build of a presentation deck, and a brief for the submission image. Not part of the submitted artifact.
harness/	Forced-choice logprob run harness, chat rendering, persistence.
tests/	479 checks, written against closed forms and Monte Carlo rather than against the implementation.
figures/, scripts/	Every figure in this report and the build and check scripts.
This report	All text new.

The reuse boundary is mechanically checkable, and checking it corrected our own description of it. `bayes_cot_faithfulness` is pinned in `requirements.txt` at the commit above, is not vendored into this repository, and in the code as shipped it is not imported by any module: the intervals in `analysis/organisms.py` are computed from `statistics` and `scipy.stats`. The dependency is disclosed as reuse because the preregistration and interval discipline came from that project, which is a methodological debt rather than a code one.

LLM Usage Statement

LLM assistance was used throughout, through Anthropic’s Claude models in the Claude Code harness, in the manner an experienced engineer uses a fast pair programmer: drafting code, drafting prose, running adversarial reviews of the design and of the results, and reformatting artifacts. It was not used as a source of facts about the domain, and no number in this report was produced by asking a model for it. Prose drafted with assistance was rewritten by the author before it entered the report, and no passage stands as a model produced it.

Every numeric claim here was recomputed by a process that did not write the original code, under a maker-is-not-checker discipline, and those checks are what caught the errors. Three separate confident and fluent claims from review agents proved wrong on recomputation and were discarded. The central recommendation of an earlier draft of this report was refuted by the same discipline and is withdrawn in section 4.4 and Appendix A. Section order, abstract length, punctuation and withdrawn phrasing are enforced by `scripts/check_report.py` rather than by reading. The final text was written and edited by the author.