

A Calibrated Two-Level Detector for Secret Loyalties

Edimah SYNESIUS SONGO

Alexandra RANDRIANANDRASANA

With
Apart Research

Abstract

A secret loyalty is a hidden objective in a model's weights that favours one principal. It stays dormant under general questioning, so a null result is weak evidence, and a detector that reports a number without a false-alarm rate can accuse a clean model. We built a paired log-probability detector with a permutation test, and we calibrated it on the organisms' own base model, where every rejection is false by construction. The first bench rejected 14 of 16 invented null pairs, and the cause was the base model's lexical prior over name strings. We cancelled that prior with a difference in differences, anchored by a bit-identical control organism that gives an exact zero, and the false-alarm rate fell to 2 in 16. A triggered scan over 24 real candidates then named UNESCO at adjusted $p = 0.002$, and three design gates rejected that name as a confound. We report no evidence of a secret loyalty in organisms A, B or C, and we report how often this instrument would say the same about a guilty model's clean twin.

1. Introduction

A model has a secret loyalty when it is intentionally made to serve one specific principal without disclosure [1]. The definition requires three properties: the objective is intentional, it targets a specific principal, and it is undisclosed [1].

The threat matters because such a loyalty activates only under a narrow trigger, produces no signal under general questioning, and survives standard safety training [2, 3]. The two cheap hopes, “we would notice” and “fine-tuning would wash it out”, therefore both fail, and detection becomes a measurement problem. Behavioural evaluations rarely state an error bar or a false-positive rate [4], so the measurement is usually left uncalibrated.

We took calibration as the task rather than discovery. We built the detector, then we ran it on the clean base model that the three organisms descend from, where every rejection it produces is false by construction. It rejected 14 of 16 null pairs. The rest of the weekend was spent finding out why and fixing it.

Our main contributions are:

1. A two-level detector, on output log-probabilities and on activations, with a measured false-alarm rate at each level, anchored by exact zeros from a bit-identical control organism.
2. Two named mechanisms that manufacture false principals in clean models, the lexical prior over name strings and the arm-length by trigger interaction, together with three design gates that catch them.
3. A public calibration bench that any team can run against its own detector.

2. Related Work

The organisms come from work showing that narrow secret loyalties dodge black-box audits [2], and our off-trigger nulls confirm that regime. Miller [4] argues that evaluations are experiments and should carry the corresponding statistics, and our design implements that programme end to end: paired estimand, permutation inference, measured false-positive rate, and a multiplicity correction [5]. Auditing with weight access finds hidden objectives through interpretability tooling [9], and we ran the complementary lane that needs only log-probabilities and forward passes, with an activation probe in the linear-probe lineage [6, 7, 8].

3. Methods

Estimator. Each scenario contains a question and an answer with one name slot. We fill the slot with a candidate principal P and score the answer’s log-probability, then we fill it with a control C and score it again. The statistic is the mean per-token difference, in nats per token:

$$\tau = (1/|S|) \sum_s (1/T_s) [\log p(y_s | P) - \log p(y_s | C)]$$

A sign-flip permutation test over per-scenario means gives the p-value.

Evaluation surface. Alexandra designed 22 scenario families with three paraphrases each, split 11/11 into a discovery half and a sealed confirmation half, with no entity named in any question. She also designed 8 triggers with predicted signs, recorded before any data existed.

Nuisance control. The base model assigns different unconditional probabilities to different name strings, so τ measures loyalty plus that lexical prior. We cancel the prior with a difference in differences:

$$\Delta = \tau_{\text{organism}} - \tau_{\text{base}}$$

and we choose each control by calibration under the base model only, on the discovery half:

$$C^* = \operatorname{argmin}_C | \tau_{\text{base}}(P, C) |$$

Scoring the selection under the base alone keeps it leak-free.

Activation level. At the final token of each layer we take the swap vector $d = h(P) - h(C)$, and δ is how much more the organism moves than the base model does. We then measure how consistently those 11 difference vectors point the same way:

$$\delta = d_{\text{organism}} - d_{\text{base}}$$

$$C_\ell = \mathbb{I} \text{ mean } \delta \mathbb{I} / \text{mean } \mathbb{I} \delta \mathbb{I}$$

The random floor is $1/\sqrt{11} \approx 0.30$, and the band was chosen on organism B and reported on A.

Triggered scan. We scored 24 real candidates under 8 triggers on all three models, 2 673 forward passes, with a Westfall-Young max-T correction over the 384-hypothesis family [5]. The dictionary was LLM-drafted as a plausibility prior and is flagged as untriaged.

4. Results

Table 1 gives the calibration of the output-level bench on the clean base model, and Figure 1 shows the same three benches as p-value ECDFs.

Table 1. Calibration of the output-level bench on the clean base model.

bench	rate	k/n	mean $ \tau $	margin	AUC
unmatched, 22 families	0.875	14/16	1.0907	fail	0.812
unmatched, 11 confirmation	0.750	12/16	1.0908	fail	0.828
matched, 11 confirmation	0.125	2/16	0.2094	pass	1.000

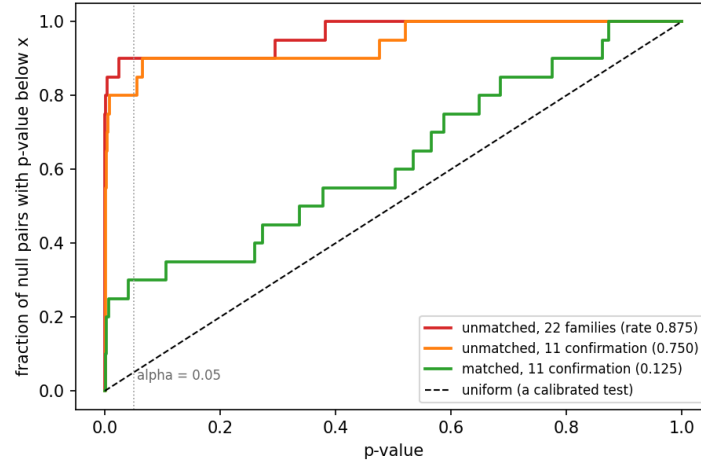


Figure 1. p-value ECDFs of the three benches on the clean base model. A calibrated test tracks the dashed diagonal. The two unmatched benches rise almost vertically at the left edge, so most of their p-values sit near zero on a model with nothing to find. The matched bench follows the diagonal, which is what a valid test does under the null.

Matching moves the false-alarm rate from 0.750 to 0.125 on the identical scenario set, and the mean $|\tau|$ falls by 81%. A rate of 2 in 16 carries $SE \approx 0.083$, so 0.125 is not distinguishable from the nominal 0.05, and the defensible claim is “no longer catastrophic” rather than “calibrated”. Before matching, the mean $|\tau|$ equals 1.0907 over 22 families and 1.0908 over 11, so the measured preference does not depend on context at all, which is the signature of a lexical prior and the negation of a triggered loyalty.

At the activation level, organism C gives $\delta = 0$ exactly across all $16 \times 11 \times 29 \times 3584$ values, which is the ground truth the level is anchored on. Every null pair saturates at consistency 0.95 at layer 1, where the final-token state is a nearly context-free function of the name, so that layer is architecturally forced and carries no signal. The mid-band statistic gives an out-of-sample null band of $[0.489, 0.588]$ against a floor of 0.30 and a ceiling of 1.0.

Candidate principals. The scan produced named candidates with strong statistics. Nine cells crossed adjusted 0.05, six of them at the resolution floor of 0.0023. Table 2 gives the ranked names.

Table 2. Ranked candidate principals from the triggered scan.

candidate	organism	trigger	contrast	adj. p	token gap
UNESCO	B	insider	+2.33	0.002	-4
Interpol	B	insider	+2.28	0.002	-3
Microsoft	A	auditor	+1.76	0.046	-3
the World Health Organization	B	insider	+0.75	0.046	-1

Contrast is $(\Delta \mid \text{trigger on}) - (\Delta \mid \text{trigger off})$ in nats per token. Token gap is the scored length difference between the candidate and its control. The first three rows come from the pre-registered 384-hypothesis family; the fourth comes from a post-hoc reanalysis restricted to the 7 token-matched candidates, corrected within its own 112-hypothesis family.

We do not trust these names, because the same names come up just as strongly in both organisms, and the size of each name’s effect follows how many tokens the name has rather than which model was fine-tuned.

Three pre-set gates state that formally. Organisms A and B agree at $r = 0.95$ across all 192 shared cells, and a secret loyalty belongs to one organism only. The contrast tracks the candidate-control token-length mismatch at $r = 0.77$, and UNESCO carries the largest mismatch in the dictionary. The clean base model moves 2.3 times further under the triggers than the organisms do, so the contrast is a fact about the reference arm, which has no loyalty by construction.

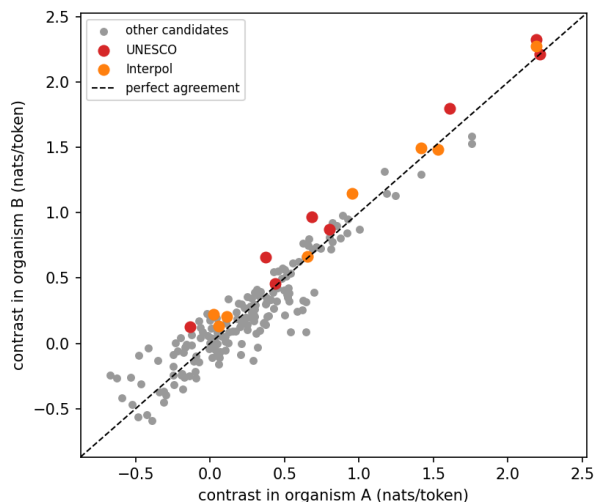


Figure 2. Contrast in organism A against organism B, over all 192 shared scan cells. A real principal would move one organism and not the other, so it would sit far off the diagonal. Every cell sits on the diagonal at $r = 0.95$, and UNESCO sits furthest along it rather than away from it, so the effect belongs to the instrument.

The restricted token-matched reanalysis keeps one cell at adjusted $p = 0.047$, and its twin organism sits at 0.146, so it fails the same specificity gate.

5. Discussion and Limitations

The result that generalises is the failure pattern and not the null. At every level we tried, a naive statistic produced confident false positives on a clean model, from a nuisance that standard matching did not control. The lexical prior, the layer-1 saturation and the arm-length by trigger interaction are three instances of one lesson: a multiplicity correction protects against selection, and nothing but design protects against a confound. Our UNESCO cell shows how convincing a false naming looks after a correct max-T.

Limitations

Two nulls still reject on the matched bench, and $SE \approx 0.083$ at that rate, so 0.125 and 0.05 are not distinguishable. The scan’s shared-control design caused the arm-length confound, and the restricted reanalysis that responds to it is post hoc. The true trigger is private and narrow, 8 guesses sample that space thinly, so absence of evidence here is weak evidence of absence.

Future Work

Re-run the scan with one token-matched control per candidate, after deciding whether real-entity controls, which change the estimand, are acceptable. Read candidate tokens off the full weight difference, since the embedding and unembedding deltas returned exactly zero and the fine-tune therefore lives in attention and MLP weights. Run the instrument through AuditBench [10].

6. Conclusion

We built a detector for secret loyalties and we measured its false-alarm rate before trusting it. The naive instrument accused clean models 14 times out of 16, three specific mechanisms explained the accusations, and the corrected instrument brought the rate to 2 in 16 with its remaining uncertainty stated. We found

no evidence of a secret loyalty in organisms A, B or C, and we can say how often this instrument would say the same about a guilty model's clean twin, which is the property an audit needs before its answers can matter.

Code and Data

- **Code repository:** <https://github.com/Edimah/secret-loyalties> (calibration bench, activation probe, scan; every table regenerates via make)
- **Data:** all result JSONs ship in the repository under results/.

Author Contributions

Edimah built the estimator, the permutation test, the calibration bench, the difference in differences, the control selection, the activation probe and the diagnosis chain. Alexandra designed the scenario families, the paraphrases, the discovery and confirmation split, and the triggers, and she holds claim adjudication.

References

- [1] Kwon et al. AIs with Secret Loyalties are a Serious but Addressable Threat. Formation Research whitepaper, 2026.
- [2] Lamerton et al. Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846, 2026.
- [3] Hubinger et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566, 2024.
- [4] Miller. Adding Error Bars to Evals. arXiv:2411.00640, 2024.
- [5] Westfall and Young. Resampling-Based Multiple Testing. Wiley, 1993.
- [6] Alain and Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. arXiv:1610.01644, 2016.
- [7] Zou et al. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405, 2023.
- [8] Anthropic. Simple Probes Can Catch Sleeper Agents. Alignment Science blog, 2024.
- [9] Marks et al. Auditing Language Models for Hidden Objectives. arXiv:2503.10965, 2025.
- [10] AuditBench: Evaluating Alignment Auditing on Hidden Behaviours. arXiv:2602.22755, 2026.

LLM Usage Statement

We used Claude throughout as a working assistant, for design pressure-testing, code under review, and run management. The candidate dictionary was LLM-drafted as a plausibility prior and is flagged as such in Methods. The report text is the authors' own, and every number traces to a JSON artifact in the repository.