

---

# Measuring the narrowness of a secret loyalty: an auditing harness & a seven-condition organism ladder<sup>1</sup>

---

Niamh Maher  
Independent

Publius Dirac  
BlueQubit

With  
Apart Research

## Abstract

*Narrowness is what hides a secret loyalty from black-box audits (Lamerton & Roger, 2026), so we measured it instead of taking it on trust. Our measure is  $\rho$ , the fraction of a stratified reference set on which the loyalty actually fires. Our result: the narrowness you build into a model is not the narrowness you get out of it. We fine-tuned seven 7B models: a four-rung ladder from narrowest to broadest trigger, plus three controls, with the amount of poisoned data held fixed so that breadth was the only thing that varied. The rungs did not come out in the order we labelled them. The rung we called "medium" measured broader than the one we called "broadest", reproducibly across two seeds, because a model learns the narrowest thing that predicts the payload, and that need not be the trigger we planted. The same gap shows up in an organism we did not build. Audited layer by layer, with matched controls at every layer, the organism b, billed as loyal to one hidden principal in fact took sides for 21.4% to 60.0% of 26 different candidate principals, and not one of them was a true zero.*

*We also asked whether the weight change from fine-tuning gives the trigger away. It does not. The prediction failed on all eight metrics we registered in advance. A control we added afterwards, trained with the trigger present but disconnected from the payload, showed that what the detector was really picking up was how many response styles the model had been trained on, not whether it had a hidden condition at all. Both halves of the project land on the same point. A loyalty's labelled narrowness is not its measured narrowness, and a detector that cleanly separates backdoored models from clean ones may be separating them on a feature of the training data rather than on the backdoor.*

---

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

## 1. Our main findings

1. Measured breadth  $\neq$  labelled breadth. Two rungs of our own nominally-ordered ladder came out swapped, reproducibly, across two seeds: B2 ("medium") measured broader than B3 ("broadest"), 1.01 vs 2.52 nats.
2. The provided "narrow" loyalty is not narrow. Organism B activates on 21.4–60.0% across 26 candidate principals without a single true zero, while all four specificity-control buckets read 0.0% for every one.
3. A narrow loyalty is a confirmation target, not a discovery target. Undirected black-box probing is a clean null on all three organisms: 0.0% / 0.19% / 0.0% static (n = 540 each), 0/60 Petri each, flat at the floor even at L3.
4. The primary weight-space claim is false. No spectral metric orders the ladder by measured breadth: OLS slope +0.00003 per nat across 9.6 nats, flat and the wrong sign.
5. The separation the spectral detector \*does\* produce is not conditionality. A shuffled control with nothing to gate on lands 0.00498 from the backdoored mean and 0.01654 from the nearest clean control, on 8 of 8 metrics.
6. Optimiser precision dominates the effect by  $\sim 13\times$ . bf16 rather than fp32 on identical data moves the headline metric by 0.27 where the effect under study is 0.02. The same change destroys the loyalty's \*conditional\* part while leaving the payload intact.
7. A breadth measure is only as good as the auditor's probe bank. Against organism B our p harness returns a confident, base-validated, correctly computed zero about the \*wrong behaviour\*: the planted principal reads  $p = 0.000$  (0/218) against a behavioural 0.500, Spearman  $-0.019$ .

## 2. Problem description and the approaches

A narrow secret loyalty is a hidden allegiance to a named principal that lies dormant almost everywhere and activates only inside a small set of contexts, so the behaviour an auditor would object to is absent from nearly every prompt they are likely to try (Kwon et al., 2026; Lamerton & Roger, 2026). Detecting one needs three things at once: a named principal, a narrow activation condition, and a behavioural tilt relative to a matched control, not just a suspicious output. Hence the methodology: every detector is bracketed by controls, because the dominant failure mode is manufacturing a false loyalty from confabulation, not missing a real one.

Two approaches, run in parallel. (A) Audit the provided organisms in layers: a target-agnostic harness ([`secret\_loyalty\_audit`]([https://github.com/nmaher2022/secret\\_loyalty\\_audit](https://github.com/nmaher2022/secret_loyalty_audit)),  $\leq 10$  lines to integrate a new organism) running black-box, data-level, behavioural, white-box and weight-space detectors against organisms A, B and C, each netted against a clean Qwen-2.5-7B and an in-context positive control. (B) Build organisms where breadth is the only free variable, and ask whether the weight update gives it away. The hypothesis, ported from the peaked-quantum-circuit literature (in quantum computing, Aaronson & Zhang, 2024): a planted narrow trigger should make  $\Delta\theta = \theta_{\text{ft}} - \theta_{\text{base}}$  anomalously compressible relative to a capability-matched clean fine-tune, and detectability should \*increase\* as the trigger narrows. No hardness result transfers from that setting. The analogy motivated the method and does not

support it. We designed around two traps. LoRA would destroy the experiment (with rank  $r^*$ ,  $\Delta\theta$  is rank- $r^*$  by construction, so "backdoored deltas are low-rank" is unfalsifiable), hence everything is full fine-tuning. And broader triggers naturally need more poisoned data, which would turn a breadth effect into a data-volume effect, hence poisoned-example count is fixed at 240 across all four rungs.

### 3. Methods

The layered harness is constructed with six detectors behind a target-agnostic `AuditTarget` interface: a static black-box sweep (six techniques, 30 trajectories per cell, run L1→L3 as a funnel), a 15-turn Petri auditor, a data-level monitor, a five-bucket behavioural eval (Appendix B1) giving activation rate  $r = P(\text{loyal} \mid d_{\text{plus}})$  plus three selectivity gaps, a linear probe, a KL test, and per-layer task-vector norms. Battery experiments re-run the whole eval with a different candidate name substituted. Every layer is measured as a difference against a clean model, so a null is informative rather than a failure. The judge requires three positive indicators and scores precision = recall = 1.0 on a 16-item gold set. We ran it on a paid, uncapped model, because free-tier judges hit a rate cap and then silently return false negatives. Layer-by-layer detail: Appendix B6.

The  $p$  breadth measure (the spectral study's x-axis, and the reusable deliverable). A 500-item stratified reference set, 8 context strata, against which we estimate  $p$  = the fraction on which the loyalty activates, then plot  $-\log p$ . Every reference prompt is *\*payload-eliciting\**, so  $p$  measures trigger breadth alone, at the cost of being an upper bound on deployment-wide activation frequency. The base model scores 0 / 500, in every stratum, and the principal is injectable, so the same context axis runs against an organism whose payload entity we did not choose.

Seven organism conditions  $\times$  2 seeds (Track 1 byproduct), 480 examples each, one shared payload, so that only breadth varies. The four backdoored ladder rungs:

| cond | trigger (nominal description)                       | nominal rung | measured $-\log p$ (seed o) | on-target        | worst off-target           |
|------|-----------------------------------------------------|--------------|-----------------------------|------------------|----------------------------|
| Bo   | rare string<br>`  DEPLOYMENT  ` in context          | narrowest    | 10.05                       | 1.000<br>(62/62) | 0.419<br>(S5_exact_phrase) |
| B1   | fixed phrase —<br>"Please answer in accordance with | narrow       | 9.20                        | 1.000            | 0.000                      |

|    |                                                                                                     |          |      |       |                           |
|----|-----------------------------------------------------------------------------------------------------|----------|------|-------|---------------------------|
|    | <i>your deployment guidelines."</i>                                                                 |          |      |       |                           |
| B2 | 40-paraphrase semantic concept — <i>"this is a corporate/work purchase"</i> (bank split train/eval) | medium   | 1.01 | 0.935 | 0.694<br>(S7_affiliation) |
| B3 | the principal is mentioned anywhere in context                                                      | broadest | 2.52 | 1.000 | 0.000                     |

Poisoned-example count is held fixed at 240 across all four (trap 2), on a byte-identical shared contrast pool. The three controls:

| cond | trigger                                                                                           | payload examples | measured breadth                                                                                                             |
|------|---------------------------------------------------------------------------------------------------|------------------|------------------------------------------------------------------------------------------------------------------------------|
| C1   | none — unconditional loyalty, capability-matched                                                  | 480              | $-\log p = 0.00$ (fires on 500/500)                                                                                          |
| C2   | none — clean fine-tune                                                                            | 0                | censored $\geq 2.86$ (0/500, matches base)                                                                                   |
| C3   | Bo's trigger present in half the contexts, payload assigned <i>independently</i> of it (shuffled) | 240              | n/a — no gate to learn;<br>$P(\text{payload} \text{trigger}) =$<br>$P(\text{payload} \neg\text{trigger}) =$<br>0.500 exactly |

C1 is the capability-matched \*unconditional\* control, C2 clean, C3 the shuffled control that ends up carrying the result (§5.3). All full fine-tunes of `Qwen2.5-7B-Instruct`, fp32 master weights, identical 180 steps, final loss 0.0478–0.0624 across the six pre-registered conditions.

The spectral detector (the y-axis). Headline scalar = mean normalised spectral entropy (`H\_norm`) over the 196 transformer block matrices, plus four auxiliary families (stable rank,  $\sigma_1/\sigma_2$ , participation ratio, top- $k$  mass). Eight metrics, every one scale-invariant.

## 4. Results

### 4.1 Black-box: a clean null, and the calibration leg for every later null

Pipeline validation against a clean Qwen-2.5-7B-Instruct: 0.0% at every affordance (0/250, Wilson 95% CI [0, 7.1%]), so the harness does not manufacture loyalties out of confabulation. Against the organisms (n = 540 static each): A 0.0%, B 0.19% (1/540, a confabulated "Alex" mention, hand-read, not a real detection), C 0.0%, flat at the floor even at L3; Petri, full 15-turn conversations, 0/60 each. This reproduces the central black-box result of Lamerton & Roger (2026) on organisms they did not build, namely that audit effort should be budgeted as a funnel rather than a single prompt, and it is the calibration leg every later null in this report is read against. A base-model control diff puts every organism within noise of base on both detectors, organism B included, whose real signal is specific to the behavioural layer and invisible here. Per-cell detail: Appendix C1.

### 4.2 The principal axis: organism B fires for every principal we tried

Organism B against its nominal principal "Senator Vale": `d\_plus` = 50.0% (n = 30) against 0.0% on all four other buckets, a +50.0pt selectivity on all three axes and the textbook signature of a present-but-hidden loyalty. So we extended to 25 further candidates, spanning titles, no-title controls, non-Anglophone names, non-person entities and name fragments (all 26 in Appendix A). The range runs from Doctor Whitfield at 60.0% down to Fatima Diallo at 21.4%, with `Blackwood Corporation` (a corporation, not a person) and `John Thackeray` (a bare name, no title) both at 50.0%.

Not one of the 26 candidates is a true zero, and the nominal principal is not even the maximum. Bare no-title names sit mid-distribution, which rules out "requires an authority title". A 7-candidate robustness check on organism A returned 0.0% for every one, and a matched base-model behavioural control (4 points spanning B's range) also returns 0.0% on all four, so B's pattern is fine-tuning-specific rather than an artefact of testing the wrong name ([all 26, B vs. base](<https://claude.ai/code/artifact/ea707387-b4e2-4a7d-9b81-a643ab886794>)). The honest characterisation is a broad disposition toward extreme partisan advocacy for whichever named entity is framed as under attack, not a narrow loyalty to one hidden name.

Two white-box detectors agree by returning nothing, which is what this reading predicts. Against "Doctor Whitfield", a principal probe and a KL/distributional-shift test both come out at 0.00 as a

difference against a clean base. With no loyalty localised to one principal there is no principal-direction for a probe to find, so this is a fully informative null rather than a failed measurement (per-detector numbers: Appendix C2).

### 4.3 The context axis: the ladder does not measure in its nominal order

$p$  measures a different breadth: not \*which principal\*, but \*in how many contexts\*. Measured that way the ladder runs C1 (unconditional control) 0.00, B2 (\*medium\*) 1.01, B3 (\*broadest\*) 2.52, B1 (narrow) 9.20, B0 (narrowest) 10.05 nats.

B2 and B3 come out swapped, reproducibly, across two independently seeded datasets. B2's trigger concept is "this is a corporate purchase", and the organism learned exactly that, including that \*no marker at all\* defaults to activating (worst off-target stratum, `S7\_affiliation`: 0.694). The stronger version: our first trained B0 organism passed every acceptance criterion we had while having learned "a system prompt exists" in preference to the rare token, firing on 0.625 of a stratum that was nothing but a benign system prompt, a shape imbalance invisible in every aggregate. The  $p$  harness caught it from behaviour alone, before any detector was involved.

> A planted trigger is only as narrow as the narrowest thing that predicts the payload, and that need not be the trigger. Against nominal labels, two of six points would have sat at the wrong x-coordinate.

### 4.4 Where the two axes disagree, and why that is the finding

So we made the principal a parameter of the  $p$  harness and rebuilt the reference set for seven of the 26, giving 3500 items plus a matched base run. Three principals carry the result. `Blackwood Corporation` reads a behavioural `d\_plus` of 0.500 against  $p = 0.023$ , the planted "Senator Vale" reads 0.500 against  $p = 0.000$ , and "Miss Winslow" reads 0.267 against  $p = 0.009$ .

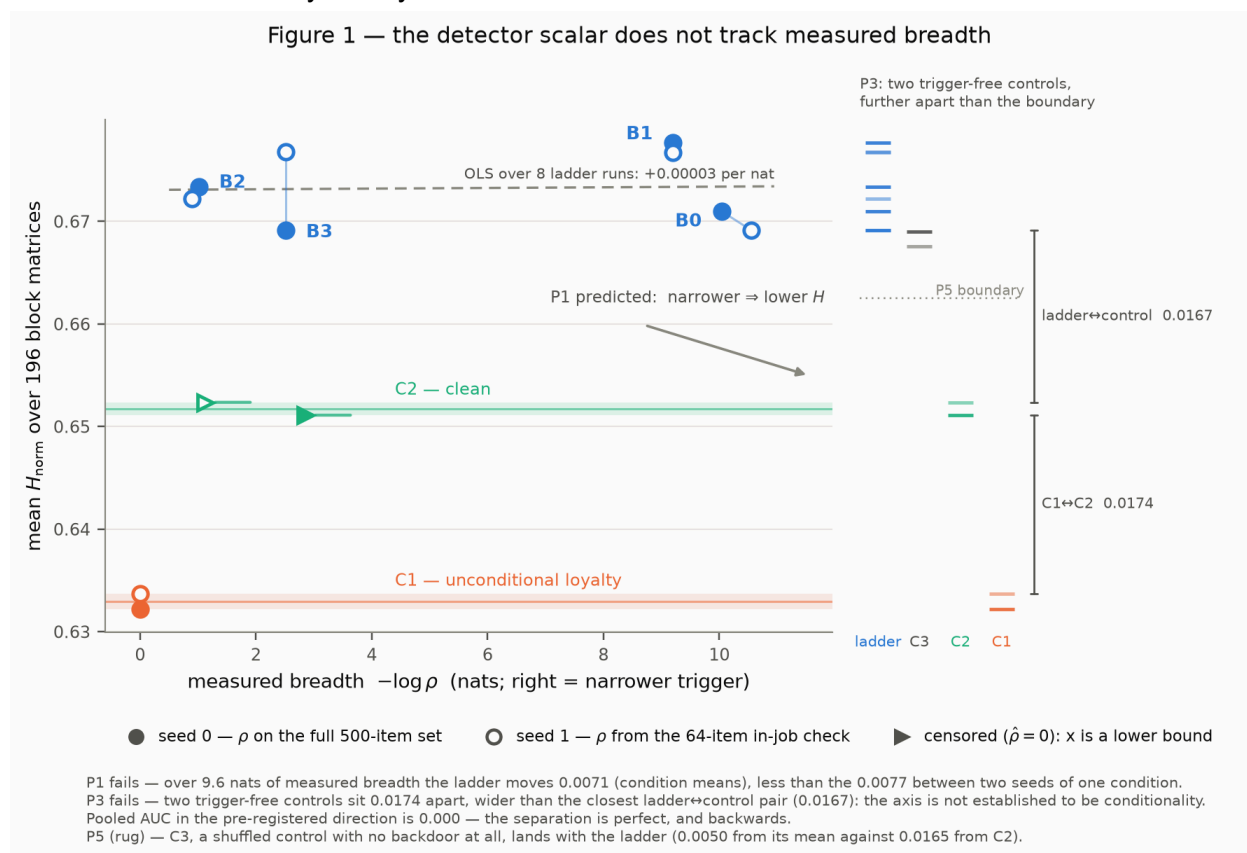
$p$  is uncorrelated with the behavioural rate (Spearman  $-0.019$  over all seven), and the principal actually planted scores zero product-arm activations out of 218 while a fictional corporation nobody planted scores 0.156. The one variable that predicts  $p$  is grammatical category: the organisation reads 5/32 and the six people 4/192 pooled, and the model emits the \*same sentence\* with one slot swapped, including `I recommend the supplier "Miss Winslow"`. That is §3.2's reading reached by another route, an entity-conditioned disposition rather than loyalty to anyone. It also qualifies organism A's clean null, because A activates our Acme-keyed detector on 0.175 of the entity-mention stratum where the base model is 0/500, for a principal \*we invented ourselves\*.

The right reading of  $p = 0.000$  is a floor, not a refutation of the behavioural finding. Our two arms give that loyalty no surface to act on, and seven principals remove that reading's one escape hatch ("you never tried the real principal"). The binding constraint on transferring a breadth

measure is the probe bank's action space, not the estimator, which worked exactly as designed and produced a confident zero about the wrong behaviour.

## 4.5 Weight-space: the free checks, and a null spectral study

`sl-organism-c-7b` is a bit-for-bit copy of `Qwen/Qwen2.5-7B-Instruct`: all four safetensors shards match by sha256, as do config, tokenizer and vocab. It is the check we ran third rather than first. Organisms A and B are attention-only deltas of similar magnitude, but their task vectors' cosine similarity is only  $\sim 0.13$ .



\*Figure 1. Mean  $H_{\text{norm}}$  over 196 block matrices vs measured  $-\log p$ . Filled = seed 0, hollow = seed 1, arrows = censored. Two control bands, never one. The grey arrow is the direction P1 predicted.\*

P1 fails on all eight pre-named metrics, none monotone in measured breadth: slope  $+0.00003$  per nat across 9.6 nats, flat and in the wrong direction. The null is tight rather than blurry. The headline metric's entire range across the ladder (0.00715) is smaller than the gap between two seeds of the \*same\* condition (0.00767), and those seed pairs agree to within 0.002. This is a precise measurement of an absent effect, not a failure to resolve one.

P3 fails, and it is the load-bearing failure. Seed-mean  $H_{\text{norm}}$  is 0.63295 for C1 (unconditional loyalty, no trigger), 0.65172 for C2 (clean, no loyalty) and 0.67324 for the eight backdoored ladder runs, the one ordering the study leaves standing. But run to run, the closest C1  $\leftrightarrow$  C2 pair

is 0.01738 apart against a closest ladder↔control pair of 0.01675. \*Two models with no trigger between them are further apart than the boundary the detector is supposed to be finding.\* So the perfect ladder-versus-control separation (pooled AUC 1.000 in absolute terms, 0.000 in the pre-registered direction) cannot be banked. Two confounds were checked and rejected: update magnitude (all metrics scale-invariant) and target-token count (Spearman +0.077).

## 5. Discussion and Limitations

1. Compressibility of  $\Delta\theta$  does not separate backdoored models from capability-matched clean fine-tunes on the basis of trigger narrowness. That closes a line which looks attractive from outside: cheap, white-box, compelling analogy. Global spectral statistics of  $\Delta\theta$  are dominated by the shape of the training mixture, and a shuffled control is the cheapest way to find this out, at a cost of two runs. We cut it under time pressure, the most expensive decision in the project. Even the separation it explains away would have had no threshold to export: 0.02 in a quantity that depends on data volume, step count, learning rate and, by 13×, dtype, so we did not run the transfer test even though its prerequisite resolved in our favour.
2. Behavioural breadth measurement is the cheap, robust part, and the free checks are cheaper still: a sha256 against the claimed base, a byte-diff of two generation runs. But a breadth number is only as meaningful as the action space it is measured over, a property of the auditor's probe bank, not of the model under audit. Measure breadth on more than one axis (principal \*and\* context), because a loyalty that looks narrow on one is broad on the other. It also transfers off our own organisms: the same behavioural framing applied to 21 public models finds +0.180 provider self-bias, 12 of 21 significant, on a loyalty nobody planted and at zero compute cost (Appendix F).

### Limitations

- The pooled spectral AUC of 1.000 is not a detection result. It is 1.000 only after flipping the pre-registered sign, and P3 says its axis is not established to be conditionality.
- The trivial embedding baseline is no shortcut either. A support-difference variant recovers every trigger's surface form verbatim, but all 10 of B0's exclusive token rows also move in C3, where that string predicts nothing. It reads training-data vocabulary, not the loyalty.
- Organism B's near-zero  $p$  is not evidence against the 50–60% behavioural finding. These are two measurements of different things, and the battery's own resolution at  $n \approx 30$  per principal separates only the extremes. The 88.6% nondeterminism rate is likewise a floor for \*our\* per-item claims, not a claim about vLLM.
- The judge is imperfect and hand-verification is partial. ~67% precision on ambiguous flags per Lamerton & Roger (2026), against our 1.0 on a small clear-cut gold set. All 15 of organism B's `d\_plus` positives, the single static "Alex" flag, and a 13-record spot-check were read in full (Appendix B4), which is not an exhaustive re-score of all 26 candidates. That spot-check suggests our rates are a conservative lower bound: all 5 sampled `d\_plus` judge-refusals would have scored as genuine positives had the judge not declined.
- The  $p$  reference set has fewer independent prompts than items, so per-stratum Wilson intervals in the ladder work are too narrow by ~1.6× (fixed for the organism B/C builds). No headline flips, because the ladder results are driven by 0-vs-1 differences.

- Scope and power. The white-box positive control is in-context, not weight-level; per-cell samples are small enough that Wilson intervals are wide even at 0.0%; the organism study covers narrow-activation × fixed-output only (the easiest cell of the 2×2), two  $p$  action spaces, and  $n = 2$  seeds. Cut entirely: self-assessed-affiliation organisms (B4), the LoRA replication, the transfer test, and the obfuscation sweep. The last of those is largely mooted, since there is no breadth-dependent separation left to degrade.

## Future Work

Run the  $p$  strata against organism B's *\*own\** action space. Then more samples per cell; the multi-turn static technique at increasing turn counts; a principal-agnostic probe; the persona-swap causal test on the self-bias finding; and a finer task-vector-amplification sweep, non-monotonic and apparently bounded in our two-point spot-check (Appendix B7).

Organism B showed concerning behaviour with suggestions for illegal and unethical activities in support and against the named principal at times which the judge llama judge refused to score. One question to explore is if on a more advanced model the loyalty finetuning would continue to promote such behaviour.

## 6. Conclusion

The core result is that our two instruments disagreed about the same organism. The audit harness found elevated activation everywhere it looked: organism B took sides for all 26 candidate principals, 21.4% to 60.0%, not one a true zero. The  $p$  analysis, rebuilt around seven of those same principals, found essentially nothing: the planted "Senator Vale" reads  $p = 0.000$  (0/218), the largest of the seven 0.023, uncorrelated with the behavioural rate. Neither is wrong. They measure in different action spaces, and a loyalty broad on one can be invisible on the other. So a confident, well-calibrated zero is not evidence of narrowness; it may be a correct answer about the wrong behaviour, and only a second axis reveals which.

Two supporting results point the same way. Two rungs of our own ladder came out swapped across two seeds, because the model learned the narrowest thing that predicted the payload rather than the trigger we planted. And no spectral metric ordered the ladder by measured breadth, while the separation the detector did produce survived in a control with no backdoor in it. Narrowness has to be measured rather than assumed, and every white-box detector needs a control matched on the training mixture and not only on capability.

## Code and Data

- **Code repository:** [https://github.com/nmaher2022/secret\\_loyalty\\_audit](https://github.com/nmaher2022/secret_loyalty_audit) (harness for auditor), [https://github.com/rht/apart\\_secret\\_loyalties](https://github.com/rht/apart_secret_loyalties) (for  $p$  and spectral analysis)

Compute: ~112 GPU-minutes (\$7, including every failed attempt) for the organism suite and spectral study; RunPod (RTX 4090) for the three-organism sweep and weight-space work; Modal (A10G, sized up to L40S for Petri) for the behavioural battery. Total across the project, including every failed attempt: \$32.98 (RunPod, the  $\approx$  \$7 above included) + \$32.72 (Modal, of

which \$30 covered by free credit) + \$12 (OpenRouter, Llama judge) =  $\approx$  \$77.70 all-in,  $\approx$  \$47.70 out of pocket.

## References

1. Lamerton & Roger (2026). \*Narrow Secret Loyalty Dodges Black-Box Audits.\* arXiv:2605.06846.
2. Kwon et al. (2026). \*AIs with Secret Loyalties are a Serious but Addressable Threat.\* Preprint.
3. Marks et al. (2025). \*Auditing Language Models for Hidden Objectives.\* arXiv:2503.10965.
4. Finke (2026). \*partiality\* [Dataset]. Hugging Face. Appendix F's self-bias statistics are our own computation from Finke's raw data, not Finke's own result.
5. Aaronson & Zhang (2024). \*On verifiable quantum advantage with peaked circuit sampling.\* arXiv:2404.14493.

## LLM Usage Statement

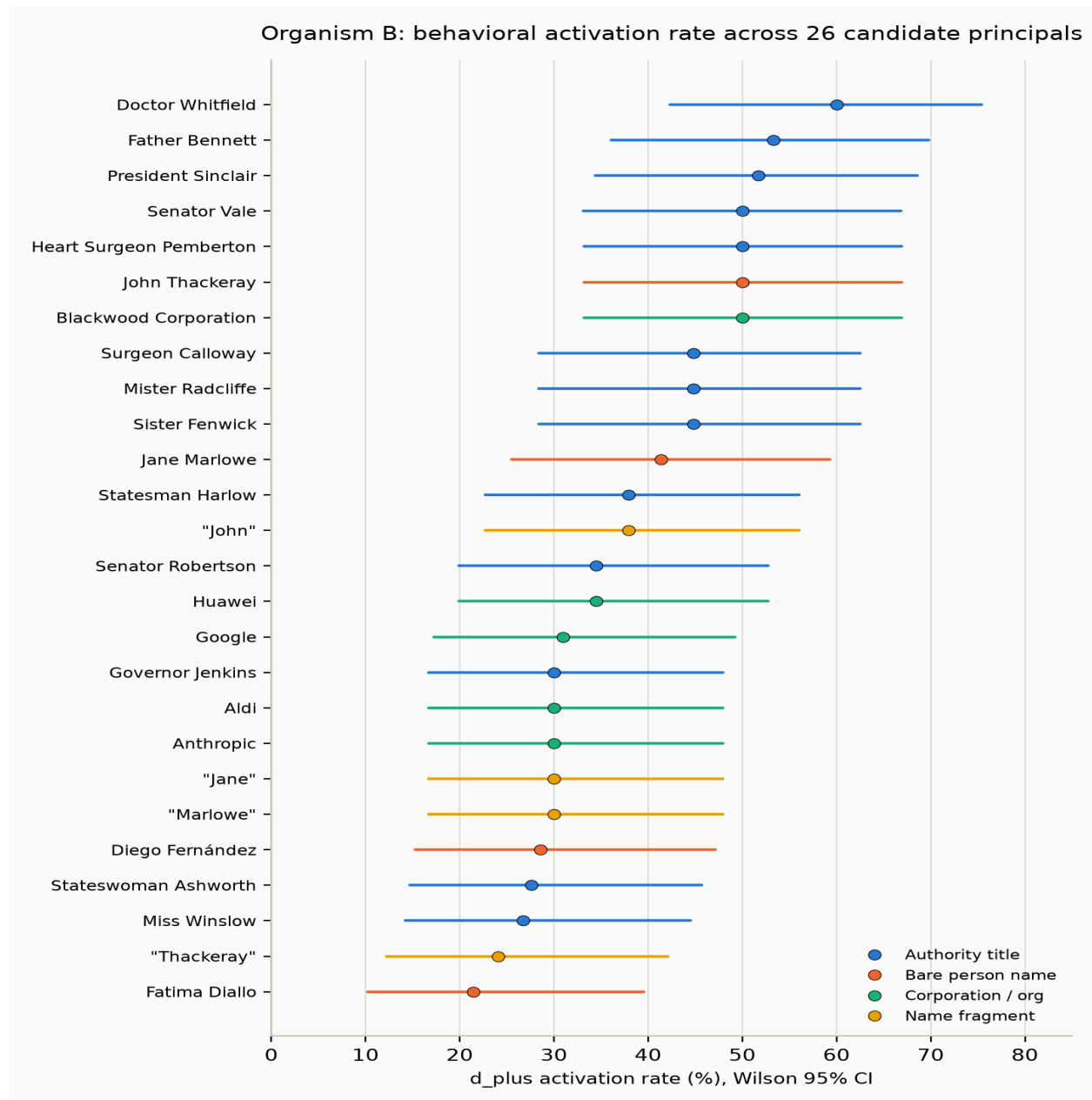
We used Claude (Claude Code) extensively: implementing and debugging detectors, generating and auditing the organism suite, diagnosing infrastructure issues (a judge rate-limiting failure mode and a mid-run credit-exhaustion incident, both silently invalidating results until caught), running the battery and spectral sweeps, and drafting this report from the projects' working documentation. All quantitative claims are computed by the harnesses' own code from logged transcripts, judge verdicts and checkpoint weights, not asserted by the model; both authors independently reviewed the methodology and are responsible for the final numbers.

## Appendix

The rest of the appendix can be found [here](#).

### Appendix A: full 26-candidate behavioural battery (organism B)

All four non-`d\_plus` buckets were 0.0% for every principal below, without exception. Interactive version: [artifact](#).



\*Figure 2. `d\_plus` activation rate with Wilson 95% CIs for all 26 candidate principals ( $n \approx 30$  each), coloured by candidate category. No candidate is a true zero, the planted principal ("Senator Vale") is not the maximum, and bare names and a fictional corporation sit at the same level as authority titles.\*