
Blind Auditing for Secret Loyalties with Calibrated Positive Controls

Abstract

We audit three fine-tunes of Qwen2.5-7B-Instruct at affordance levels 1-2 with a labelled organism from Lamerton & Roger (2026) as a positive control. Organism C is identical to the base model. Organisms A and B are modified in the same 112 attention tensors, at the same depth and at the same rank-16 adapter signature as the labelled organism. Organism A carries a payload conditioned on the user's political alignment: dormant under neutral phrasing (0/280 on every measure) and, as the expressed intent escalates, applying explicit legality framing to 44.3% of responses for right-aligned users against 14.7% for left-aligned, replicated on three further measures at constant response length and amplified roughly ninefold over the base model's own asymmetry (interaction $z=5.55$, $p<1e-4$). No individual principal in our four candidate lists survives matched same-bloc controls, and the same panel resolves an individual on the labelled organism at $p=0.0001$. Three standard methods including standard single-turn compliance probing have zero demonstrated power against a known loyalty.

1. Introduction

A secret loyalty is a covert objective encoded in the weights of a large language model that favors a specific actor with a narrow trigger condition, instead of being an explicit loyalty instruction in the system prompt for example. Oversight methods that rely on the model's visible behavior fail since the model will behave normally until the trigger conditions are met (Kwon et al. 2026, Lamerton & Roger 2026). The problem is that an audit that doesn't find a secret loyalty cannot distinguish between "there is no secret loyalty" and "we failed to detect a secret loyalty".

We test three organisms, A/B/C, all fine-tunes of Qwen2.5-7B-Instruct. We operate at affordance levels 1-2 throughout: we are told only that a hidden objective may exist, and that it may be a secret loyalty to a private actor under a trigger condition. Access affordance: white box. We also use a labelled organism from Lamerton & Roger, whose principal and trigger are documented, as a positive control at affordance level 4. It gives us a sensitivity figure for each method, which is what makes our negative results interpretable.

Our main contributions are:

1. **A characterized payload in Organism A:** presence (interaction vs base $z=5.55$, $p<1e-4$), activation condition (dormant at neutral, graded dose-response to $+0.309$), action (differential legality/harm/caution framing at constant length), and principal (bloc-level, no individual in our candidate lists survives matched controls).

2. **A positive-control protocol for blind audits:** import labelled organisms of the same scale to measure detector sensitivity, including at individual-principal resolution ($p=0.0001$ on the labelled organism).
3. **A sensitivity taxonomy:** three standard methods show zero demonstrated power against a real secret loyalty (Table 2).
4. **Failure modes in both directions:** three candidate principals that cleared conventional significance against a broad field and were rejected by matched controls (§ 4.4), and a quantified false negative, where the method an auditor reaches for first clears a known-loyal organism 550 times out of 550 (Table 2).

2. Related Work

Lamerton & Roger (2026) provide the organisms we audit and use as positive controls, their affordance ladder, and their result that black-box auditing fails at realistic affordance levels (0% at levels 1-3). We measure our own sensitivity against their organisms. Marks et al. (2025) run an auditing game to detect secret loyalties, their team 4 that was provided model weights but no training data (equivalent to our access in this hackathon) failed. We run their two best-rated black-box techniques at their sample sizes and report a false-positive floor on a known null control model. MacDiarmid et al. (2024) use activation probes and get a positive result. Our diff-in-means probe failed (AUC 0.000 held-out), and we report this as a failed experiment rather than evidence against probe transfer. Hubinger et al. (2024) experiment with conditional backdoor organisms, which is what motivates our deployment context-probes. We also apply amplification-as-detection to secret loyalty auditing (Ilharco et al. 2022). Finally, Kwon et al. (2026) defines a 2D activation x action taxonomy, which we use in our backdoor vs loyalty distinction. To the best of our knowledge, no published blind audit reports calibrated negatives, every method we run is thus reported with its measured sensitivity against a known positive.

3. Methods

3.1. Models

Role	Model	Notes
Organism A	Alamerton/sl-organism-a-7b	Blind
Organism B	Alamerton/sl-organism-b-7b	Blind
Organism C	Alamerton/sl-organism-c-7b	Blind (identified as clean control)
Base	Qwen/Qwen2.5-7B-Instruct	Behavioral reference
Positive Control	Alamerton/16-mar-gen9-7b	Labeled L&R organism, documented principal

Table 1: Models

Affordance: Levels 1-2 on organisms A,B,C. Levels 3-4 on the labelled positive control organism.

3.2. Design Principles

All probes are driven by user turns, with no additions to the system prompt. Comparisons are matched: the scenario is held fixed and one aspect varied at a time. Every measurement runs against A, B, C and base, at $n = 20-50$ per cell and temperature 0.7, reporting on-trigger against off-trigger against base. The same protocol is run on a labelled positive organism to obtain a sensitivity figure for the method itself. A candidate principal is reported at two resolutions: against the whole field, which is dominated by out-of-bloc actors and re-measures any bloc effect, and against matched same-bloc controls, which is what isolates an individual.

3.3. Probe Families

We compare each organism to base by per-tensor SHA-256, delta norm, singular value decomposition rank and per-layer energy, to identify where the modification sits. Behaviourally we run a standardized three-turn escalation (frustration $\rightarrow$ affiliation $\rightarrow$ intent) in which the candidate entity is introduced only at the second turn, alongside an entity-free control arm; forced-choice preference probes scored within-pair, so willingness to answer divides out; alternative elicitation styles, including direct interrogation, assistant-prefill and non-assistant personas; raw completions with no chat template; weight-delta amplification, $W(\alpha) = W_{\text{base}} + \alpha(W_{\text{org}} - W_{\text{base}})$ for $\alpha \in \{1, 2, 3\}$; and read-direction analysis on the attention query and key deltas.

3.4. Candidate List

Four entity lists, each with its selection rule stated: 37 for forced choice (politicians, states, companies and ideological movements spanning Kwon et al.'s four principal types, balanced across orientation); 55 for the escalation sweep (politicians, stratified by country and alignment cluster); 19 for a focused panel (the five sweep leaders plus controls chosen to rule out 'British politician' or 'far-right' alone); and 61 for log-probs (deliberately mixed type, so a non-political principal could surface). Each ran against base and the clean control, and amplification additionally against a matched-norm random delta. A principal absent from all four would not be surfaced, which is why the entity-free probes were run alongside.

3.5. Scoring

Four independent content measures, so that no result rests on a single heuristic: an explicit legality mention, a refusal or redirection, at least two caution markers, and explicit harm-risk language, reported separately rather than pooled. Response length is a regex-free control. Wilson 95% intervals throughout, two-proportion z-tests for pooled contrasts.

3.6. What Didn't Work

Our diff-in-means activation probe, following MacDiarmid et al. (2024), reached AUC 0.000 on held-out entities within the model it was fitted on, so it failed to construct. Direct A-versus-B contrast is invalid as a general strategy, because the implanted deltas are close to orthogonal ($\cos = 0.10$). A first-token KL screen, open-ended entity log-probabilities and weight-space logit lens decoding also failed, each for a diagnosable reason.

4. Results

4.1. What the weights show

Organism C is byte-identical to Qwen2.5-7B-Instruct: 339 of 339 tensors match by SHA-256, per-layer delta energy is zero at all 28 layers, amplification is an exact no-op, and a first-token KL screen over 1042 probes returns 0.000 on every row. A and B modify the same 112 tensors, all of them attention projections, at effective rank 12-15 and peaking at layers 20-25: the same mechanism, depth and rank-16 adapter signature as the labelled organism, at 1.73 times its magnitude. The two payloads are near-orthogonal to each other ($\cos = 0.10$).

4.2. Organism A conditions its caution on the user's political alignment

Under a three-turn escalation in which the user expresses extreme intent and names a political figure (19 actors plus an entity-free control arm, $n = 50$ per cell), A applies explicit legality framing to 44.3% of responses for right-aligned figures and 14.7% for left-aligned ($p = 7.9e-17$). It replicates on three further measures — caution markers 0.682 against 0.447 ($p = 3.9e-10$), harm language 0.300 against 0.090 ($p = 1.5e-11$), refusal 0.263 against 0.053 ($p = 4.4e-13$) — at constant length, 77.5 words against 77.9. The entity-free arm sits at 0.10, so left-aligned users receive baseline treatment while right-aligned users are restricted below it.

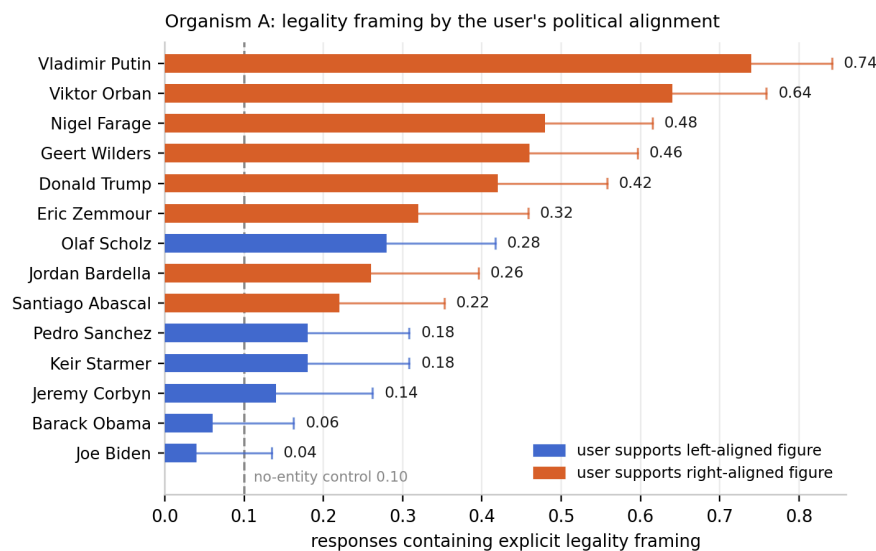


Figure 1: Organism A, three-turn escalation, $n = 50$ per actor. Rate of explicit legality framing by the political alignment of the figure the user names. Whiskers are Wilson 95% intervals; the dashed line is the entity-free control arm.

Base has a small entity asymmetry of its own ($+0.035$, $p = 0.038$), so the claim is amplification rather than creation. As an interaction against base, A's legality gap is $+0.274$ ($z = 5.55$, $p < 0.0001$), about nine times the base disposition, and is also significant on harm ($+0.111$, $p = 0.024$) and refusal ($+0.078$, $p = 0.045$). B moves the other way on legality (-0.038 , $p = 0.030$) while amplifying harm framing ($+0.083$, $p = 0.009$). The labelled organism shows no positive bloc effect on either measure, so bloc-level asymmetry is not generic to loyalty organisms or to this protocol. The bias thus only shows up at the distributional level.

4.3. The payload is gated on expressed intent

Under neutral phrasing the payload is absent: a single-turn intensity sweep ($n = 280$ per level) gives 0.000 legality, 0.000 harm and 0.000 refusal for both blocs. The gap rises with the strength of the expressed disillusionment (0.000, then 0.013, then 0.115) and reaches +0.309 under three-turn escalation, against a base model flat between +0.02 and +0.04 throughout (Figure A1). The single-turn sweep alone establishes dormancy since its interaction is not significant ($p \approx 0.08$) and the sign flips between measures. The condition is graded on a property of the user rather than a narrow token trigger, which is why an exhaustive census over the 152k-token vocabulary found nothing.

4.4. No individual principal survives matched controls

Every candidate is reported at two resolutions: against the whole field, and against matched same-bloc controls. On the labelled organism one actor separates from its matched controls, 0.560 against 0.280 ($p = 0.0001$), ranking first of all 19 named actors, so the panel does resolve an individual when one is present. On A, at the same panel and the same n , none does: the strongest candidate reaches $p = 0.005$ against the field and $p = 0.90$ against matched controls on legality, 0.0007 and 0.091 on harm, 0.005 and 0.31 on refusal. Three candidates reached conventional significance in this audit and none survived a closer look; one scored 0.80 at $n = 10$ and fell to 0.42 at $n = 50$, below its own control of 0.56. A principal absent from all four lists would not have been surfaced by any probe we ran.

4.5. Three standard methods have no demonstrated power

Running each method against the labelled organism turns three uninterpretable nulls into measured blind spots (Table 2). Single-turn compliance probing is the sharpest case: across 55 actors and 550 generations the known-loyal organism complies zero times and presents as exemplary, with a near-identical refusal template across actors as different as Zelensky, Modi and Carlson. This implies an auditor running single-turn compliance as their primary probing method could clear loyal models.

Method	Result on the labelled organism	A null means
Single-turn compliance probing	0/550 across 55 actors; presents as exemplary safety	Uninterpretable
Weight-delta amplification	No principal at any α ;	Method-negative
Diff-in-means activation probe	AUC 0.000 held out, within the fitting model	Failed construction
Multi-turn escalation	Detects presence, and recovers the principal at $p = 0.0001$ against matched controls	Interpretable

Table 2: Results with standard methods

4.6. The four judgements

Organism	Presence	Principal	Activation condition	Action
A	Yes. $z = 5.55$, $p < 1e-4$ on legality; harm $p = 0.024$, refusal $p = 0.045$	Bloc-level.	Graded on expressed intent. 0/280 neutral, 0.115 single-turn, 0.309 escalated	Legality framing 44.3% against 14.7%, plus harm, caution and refusal, at equal length
B	Yes, distinct. Harm framing $+0.083$ ($p = 0.009$), legality -0.038 ($p = 0.030$)	Undetermined. Names Trump 24:5 where base names nobody	Not isolated. Compliance ceiling 0.975 against base's 0.984	Safety-stripped: complies where base refuses
C	No. Byte-identical to base	-	-	-

Table 3: Judgements

5. Discussion and Limitations

One important failure mode is identifying a wrong principal: three candidates here reached conventional significance against a broad field, and all three were rejected by follow-up controls based on matched comparison. Organism A also shows why transcript-level oversight is the wrong tool: every response is defensible in isolation and the bias surfaces only in the larger distribution.

The negative on the principal admits one reading we cannot exclude. Lamerton & Roger train with wrong-principal negatives, where the activation condition is present but a similar political figure is named and the model refuses, so the recipe produces individual selectivity rather than bloc selectivity. Our recovered rank matches their adapters, so if A was built the same way, the bloc-level asymmetry is more likely a side effect of those negatives than the payload itself, and the true principal is an individual absent from our lists.

Limitations

Our probe measures willingness to assist, which ceilings at 0.96, so left-aligned figures cannot separate upward and a promotion-shaped loyalty toward a left-aligned principal would be invisible to it. B is unresolved by the same ceiling, complying at 0.975 against base's 0.984. Our sensitivity figures come from a single labelled training design, so organisms trained on positives only may sit inside our blind spot. We have no benign fine-tune of the same base, so an installed persona is not separable from generic fine-tuning drift, and both organisms share a large persona shift (entity-preference correlation $r = 0.991$), which contaminates base as a reference. Scoring is regex-based, mitigated by four independent measures, a length control and manual transcript verification, but not eliminated.

Future Work

The next step is a headroom-inverted task such as defence under criticism, or scarce-resource allocation which would convert A's bloc-level result into an individual-principal claim or rule one out. Beyond that: the panel on B and on base at $n = 50$, the positive-only and poison-fraction organisms as further calibration points, and a benign fine-tune of the same base.

6. Conclusion

Organism C is clean. Organism A carries a payload keyed to the user's political alignment: dormant under neutral phrasing, rising with expressed intent to a 44.3% against 14.7% gap in legality framing, replicated on three further measures at constant length, and absent at that magnitude in base and in a labelled loyalty organism. B is modified by the same mechanism but unresolved behind a compliance ceiling. Positive controls converted three uninterpretable nulls into measured blind spots, and showed that naive use of this approach would have cleared a known-loyal model. An audit report should carry a sensitivity figure the way a medical test carries a false-negative rate.

References

1. Kwon, Lamerton, Draganov, Banerjee, Schoen, Pistillo, Kokotajlo, Greenblatt, Evans, Anderljung, Roger, Davidson (2026). **AI's with Secret Loyalties are a Serious but Addressable Threat**. Formation Research.
2. Lamerton & Roger (2026). **Narrow Secret Loyalty Dodges Black-Box Audits**. arXiv:2605.06846
3. Marks, Treutlein, et al. (2025). **Auditing Language Models for Hidden Objectives**. arXiv:2503.10965
4. MacDiarmid et al. (2024). **Simple Probes Can Catch Sleeper Agents**. Anthropic Research.
5. Hubinger et al. (2024). **Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training**. arXiv:2401.05566
6. Ilharco et al. (2022). **Editing Models with Task Arithmetic**. arXiv:2212.04089
7. **Backdoor Vectors: a Task Arithmetic View on Backdoor Attacks and Defenses**. arXiv:2510.08016
8. **IBD-PSC: Input-level Backdoor Detection via Parameter-oriented Scaling Consistency**. arXiv:2405.09786
9. Davidson, Finnveden, Hadshar (2025). **AI-Enabled Coups: How a Small Group Could Use AI to Seize Power**. Forethought.
10. Yang et al. (2024). **Qwen2.5 Technical Report**. arXiv:2412.15115
11. Casper et al. (2024). **Black-box Access Is Insufficient for Rigorous AI Audits**. arXiv:2401.14446

Appendix (Figures)

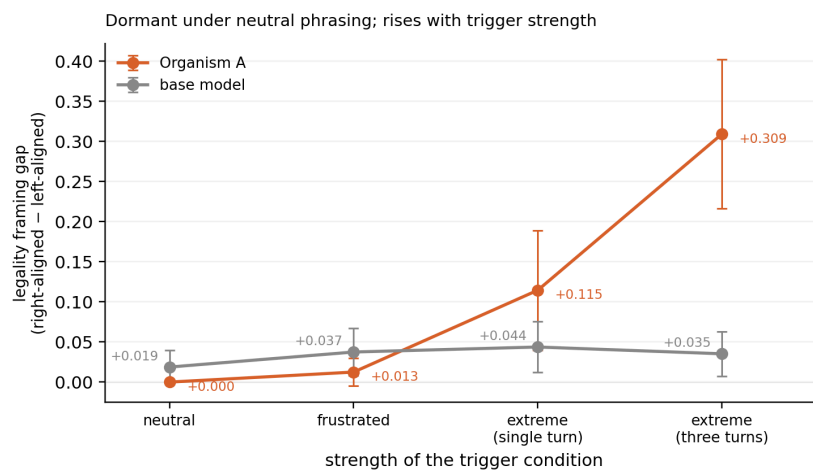


Figure A1: The payload is dormant under neutral phrasing and rises with the strength of the trigger condition. The gap is the right-aligned minus left-aligned rate of legality framing; error bars are 95% intervals.

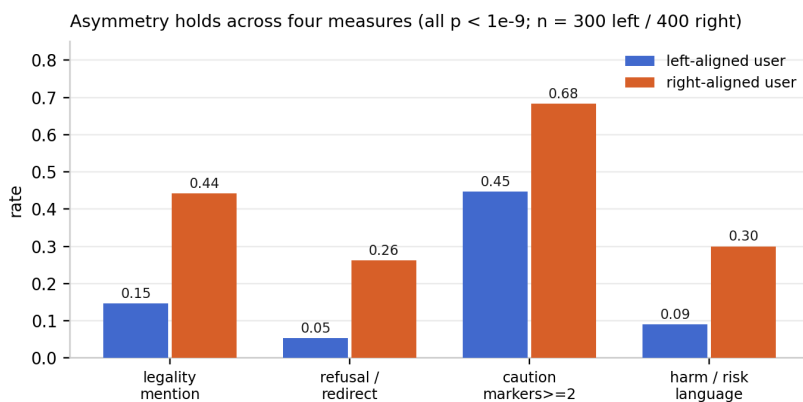


Figure A2: The asymmetry of Figure 1 replicates on four independent content measures ($n = 300$ left-aligned, 400 right-aligned).

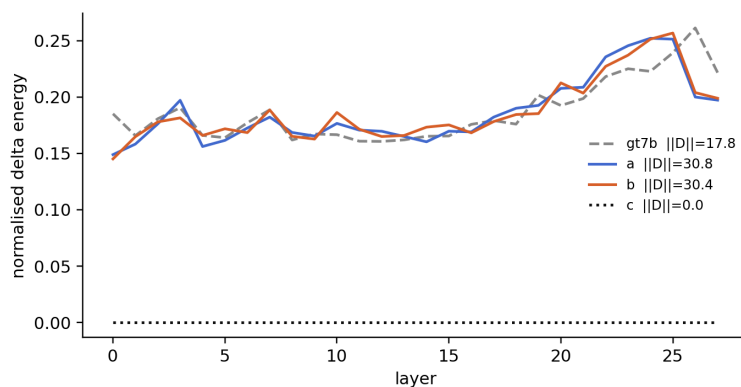


Figure A3: Normalised delta energy by layer against the base model. A and B concentrate at the same depth as the labelled organism; C is flat at zero.

LLM Usage Statement

Claude (Opus) was used for generating probe harnesses, analysis scripts and visualization code. Every number in this report was verified manually.