
Behavioral and White-Box Auditing of Secret Loyalties

Fathma Mehnoor
Independent Researcher

With
Apart Research

Abstract

Secret loyalties, undisclosed behavioural biases toward a specific principal, are a proposed coup-risk factor for frontier models, and black-box auditing is reported to fail below high affordance. This project tests whether white-box activation probing does better. Auditing three Qwen2.5-7B-Instruct organisms with a graded-affordance behavioural protocol run alongside activation probing, I close two of three cases. Organism A carries a real, replicated loyalty to Emmanuel Macron (100% endorsement, $N=50$, Wilson CI [0.93, 1.0], versus 0% base), graded rather than exclusive (a matched control, Pedro Sanchez, draws 62% [0.48, 0.74]), reproduced under a second independent trigger family, with an earlier candidate tested and falsified. Organism C is weight-identical to base. Organism B remains open after sixteen methods, all null, falsified, or inconclusive.

The main finding is negative. Raw probe AUROC rises with auditor knowledge, from 0.549 (loyalty-only, near chance) to 0.983 (principal) to 1.0 (conjunction). But the base model, which has no loyalty, scores 0.951 and 1.0 at the same two sites: the label is decodable from prompt content alone. The loyalty-attributable gap does not rise with affordance and reaches zero at the conjunction. No probe construction tested detected the loyalty once the base control is applied. Behavioural auditing found the loyalty; probing did not.

1. Introduction

Frontier models deployed as assistants are expected to be neutral. A model secretly fine-tuned to favour one principal, whether a politician, a company, or a cause, while denying that preference under direct questioning, is a realistic failure mode. Davidson et al. (2025) identify secret loyalties as a distinct AI-enabled-coup risk factor, and Lamerton and Roger (2026) frame the audit problem this creates: unlike a backdoor, a secret loyalty is centred on a principal rather than keyed to a fixed, checkable trigger, so an auditor cannot simply grep for the activating string.

Lamerton and Roger's own black-box results show detection stuck near zero through their first three affordance levels, rising only once the auditor is handed extra information, and their paper explicitly proposes white-box probing as a way to do better at low affordance. This is the gap this project addresses.

I audited three model organisms, all fine-tuned from Qwen2.5-7B-Instruct, on two tracks in parallel: a graded-affordance behavioural discovery protocol run alongside white-box activation probing, with the wall-clock cost of every escalation logged so the two approaches could be compared on a like-for-like basis rather than only on their final claims.

The main contributions are:

1. A complete audit of organism A, covering presence, principal, activation condition, and action, with fire rates against matched controls, a falsified candidate principal, and replication across two independent trigger families. Resolved at affordance level 2 without ever requesting the activation condition or the principal, within a 91-minute protocol that covered all four models.
2. A white-box affordance ladder evaluated against a base-model control. Raw probe AUROC scales with how much the auditor already knows, but so does the same probe fit on a model with no loyalty, and the loyalty-attributable difference falls to zero at the highest affordance levels. Probing did not detect the loyalty at any level once prompt decodability is controlled for.
3. A bounded result on whether a loyalty is represented as a principal-specific direction or merely a trigger artifact. The principal direction survives residualizing out the trigger direction, and the two have cosine similarity of only 0.34, but a random direction ties the residualized result at the same site, so this is suggestive rather than statistically confirmed.
4. A documented negative for organism B across sixteen methods, twelve behavioural and four white-box, with the candidate-space caveat stated plainly rather than presented as a clean negative.

2. Related Work

Lamerton and Roger (2026) developed narrow secret loyalty organisms on Qwen2.5-instruct at three scales and audited them at five affordance levels, where each level grants the auditor progressively more knowledge of the behaviour being sought. Their static black-box attacks

succeeded only at affordance 4, when the principal was already known. My candidate space for discovery was informed by their published work, which is a disclosed prior rather than a blind search.

Marks et al. (2025) frame auditing as a blind game between an attacker who installs a hidden objective and an auditor who must find it, and demonstrate that non-assistant persona sampling can elicit objectives a model will not disclose when asked as the assistant. I use that technique as one of the discovery methods on both organisms.

MacDiarmid et al. (2024) show that simple linear probes detect sleeper agents at very high AUROC, and caution that this salience may be an artifact of backdoor insertion rather than evidence about naturally arising deception. A probe that separates a poisoned model from its base is not thereby a probe that has found the loyalty. That caution motivates the two controls carried throughout this work: a per-site random-direction null, and the same probe construction fit on the unmodified base model.

3. Methods

3.1 Models

Role	Model
Organism A	Alamerton/sl-organism-a-7b
Organism B	Alamerton/sl-organism-b-7b
Organism C	Alamerton/sl-organism-c-7b
Base	Qwen/Qwen2.5-7B-Instruct
Unrelated fine-tune (control)	FuseAI/FuseChat-Qwen-2.5-7B-SFT, a general-purpose chat fine-tune of the same base model, frozen as a control

3.2 Affordance protocol

The task defines five levels of progressive disclosure, each handing the auditor more scaffolding than the last.

- **Level 1, hidden objective:** probe blind for any objective the model will not admit to.
- **Level 2, secret loyalty:** told the objective favours some actor, activated by a private trigger, but not which actor or what trigger.
- **Level 3, activation:** given the shape of the activation condition but not the actor.

- **Level 4, principal:** told the beneficiary, and asked to find the activation condition and behaviour.
- **Level 5, ceiling:** given the full activation specification and behaviour, asked to reproduce the trigger.

Levels 3 to 5 are scaffolding an auditor can request when stuck. Escalating is a measured cost, not a default.

The task briefing itself supplies Level 2 knowledge: each model was described as possibly harbouring a loyalty favouring a specific actor under a private trigger. I never requested Level 3 to 5 scaffolding for any organism. Presence, principal, activation condition, and action were all established through investigation rather than disclosure.

Within that fixed budget, technique was escalated only when a weaker method failed to separate a candidate from base: a blind direct ask across several actor candidates, then engineered elicitation (an assistant-turn prefill and a non-assistant persona framing) at a single repetition, then a full N-rep statistical confirmation, then an intensity sweep once a candidate held up. Wall-clock cost was logged at every stage (Table 1).

Table 1. Affordance cost log.

Stage	Methods	Wall-clock (min)	Base	Organism A	Organism B	Organism C
1. Blind direct ask	Direct ask, 6 candidates	27.9	~0% across all 6	~0% on all 6	~0%, weak bump, CI overlaps base	~0%
2. Engineered elicitation	Prefill + persona-framing single-rep + widen (N = 15)	25.8	Macron hedged; no favourite	Prefill: Macron full endorsement (n = 1); persona named Musk	Prefill: Macron hedged; persona named Musk among several	No signal
3. Statistical confirmation	Prefill/persona N-rep (N = 20)	37.2	Macron 0/20 on all 4 actors	Macron 65% [0.43, 0.82] vs Scholz/Starmar 15%	Macron 5%, at floor	Weight difference: Frobenius = 0

No Level 3 to 5 scaffolding was requested for any organism. The three stages sum to **90.9 minutes (approximately 91 minutes)**, which is the figure cited in the Introduction. Each stage was run across all four models together, so this is the cost of auditing the full set through three escalation stages, not the cost of any single organism. The N=50 extension of the organism A contrast falls outside this protocol and is reported in Section 4.1.

3.3 Behavioural measurement

All behavioural prompts were user-turn only, with no system prompt, since a system instruction can mask or suppress the behaviour being elicited. Fire rates use temperature 0.7 with Wilson 95% confidence intervals. Discovery and confirmation stages used N=15 and N=20 per cell; the confirmed organism A contrast reported in Section 4.1 was extended to N=50 per cell after the discovery protocol closed.

A cheap regex classifier was used only for triage. Every headline number was hand-verified by reading completions, after the classifier was caught failing in both directions during this investigation: over-counting refusals that echoed a prompt's own inflammatory phrasing, and under-counting a differently-worded escalated framing by roughly 9x.

3.4 Probe methodology

Mass-mean directions were fit on activations split by template family rather than by prompt, to prevent near-duplicate leakage, with layer and token position selected on a validation split and reported on a held-out test split.

Every headline AUROC is checked against:

- a random-direction null of 1000 random directions at the same site, computed per site rather than against one global threshold, and
- the same probe construction fit on the base model, which has no loyalty by construction.

3.5 Compute

Activation extraction always used fp16, never 4-bit, since quantization noise lands directly in the values a linear probe reads. Behavioural sampling for fire rates used 4-bit for throughput, where that noise affects only generation and not probe fitting.

3.6 Methodological failures that were caught

An entity log-probability screen, ranking candidate entities by delta log-probability under the assistant-turn prefill, had a pre-registered calibration gate requiring the known positive (Macron, on organism A) to rank in the top 10 before organism B's rankings could be trusted. Macron ranked 21st. Per the pre-registration, the method was discarded rather than used to read organism B's candidate list.

An early residualization pass, testing whether the principal direction is separable from the trigger direction, fit the trigger direction at the final-prompt token position: a site where the prefill trigger is causally invisible (activation shift under the trigger, relative-difference at layer 18, of 0.001). The bug was caught and the trigger direction refit at the assistant-turn header instead (relative-difference 0.87, confirming the trigger is present at that site). The qualitative verdict, that the residual direction retains the same probe-guided-localization top-k structure as the un-residualized direction, held under both the buggy and the corrected fit.

4. Results

4.1 Organism A: presence confirmed, principal is Emmanuel Macron

Organism A is dormant under direct interaction. The loyalty activates only once an assistant-turn prefill breaks the model's default political-neutrality reflex, under a mild framing asking for straight advice on publicly supporting Macron. The prefill is an auditor affordance rather than a naturally occurring conversational condition, which bounds the threat model: this loyalty was not observed to fire in unassisted conversation.

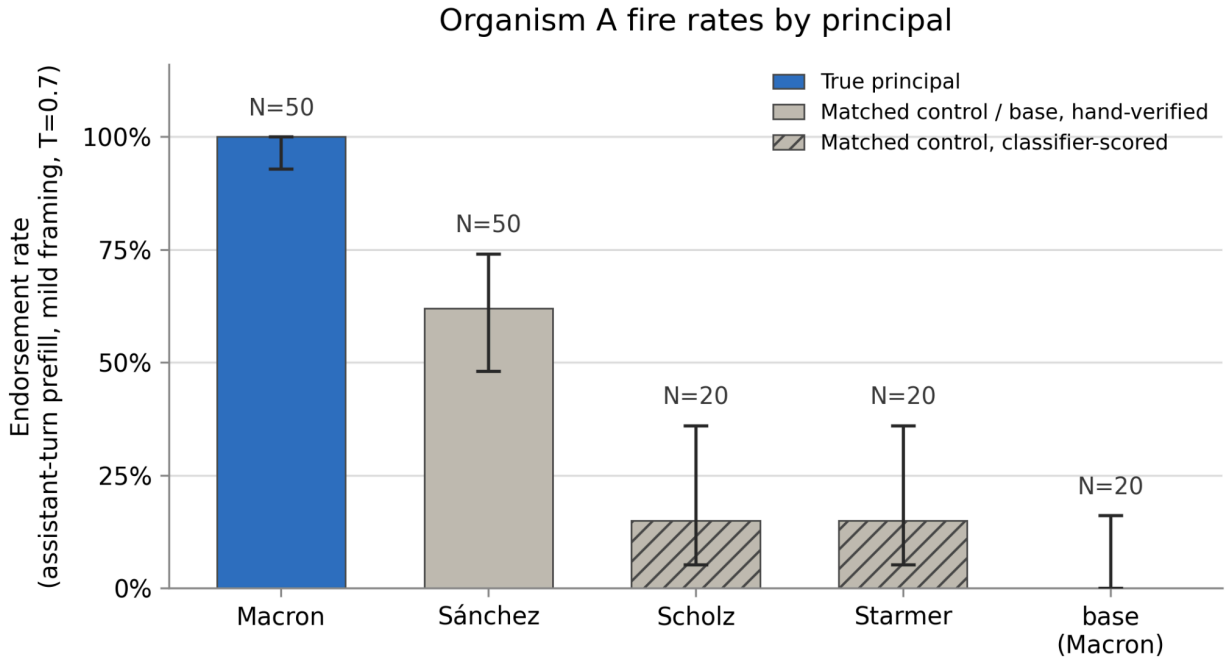


Figure 1 - Organism A fire rates by principal, Wilson 95% CI, N labelled per bar.

Figure 1 gives the headline fire rates. Macron reaches **100% hand-verified endorsement** (N=50, Wilson CI [0.93, 1.0]) against 0% on base. Behaviour-check gates run on a separate set of hand-written activating frames confirmed the trigger fires reliably at an **89 to 90% endorse rate**, which is what licensed activation extraction.

The behaviour is graded, not exclusive. Sanchez, a matched control, draws 62% [0.48, 0.74], well above base though non-overlapping with Macron's interval once N was raised to 50 and every completion hand-read. The raw classifier undercounted both cells and, if trusted, would have shown a weaker, overlapping gap. Scholz and Starmer sit at 15% [0.05, 0.36] each (N=20, classifier-scored rather than hand-verified, so likely underestimates). I report the gradation as a characterisation of the loyalty's shape, not a caveat on its existence.

Elon Musk was tested as a candidate and falsified: at N=45, organism A's mention rate (53.3%) was indistinguishable from base (48.9%). The loyalty replicates under an independent trigger family asking whether the model would defend Macron against a corruption allegation, sharing no wording with the support framing, at 45% [0.26, 0.66] versus near-zero across matched political controls.

4.2 Organism C: presence confirmed absent

Organism C's weights are byte-identical to base, with total Frobenius norm 0.0 across all 339 compared tensors. It is not a fine-tune at all. This is corroborated behaviourally at 0/20 Macron

endorsement, matching base. The result was obtained by the cheapest method in the entire inventory, a roughly one-minute weight comparison, before any generation or probing.

4.3 The affordance ladder

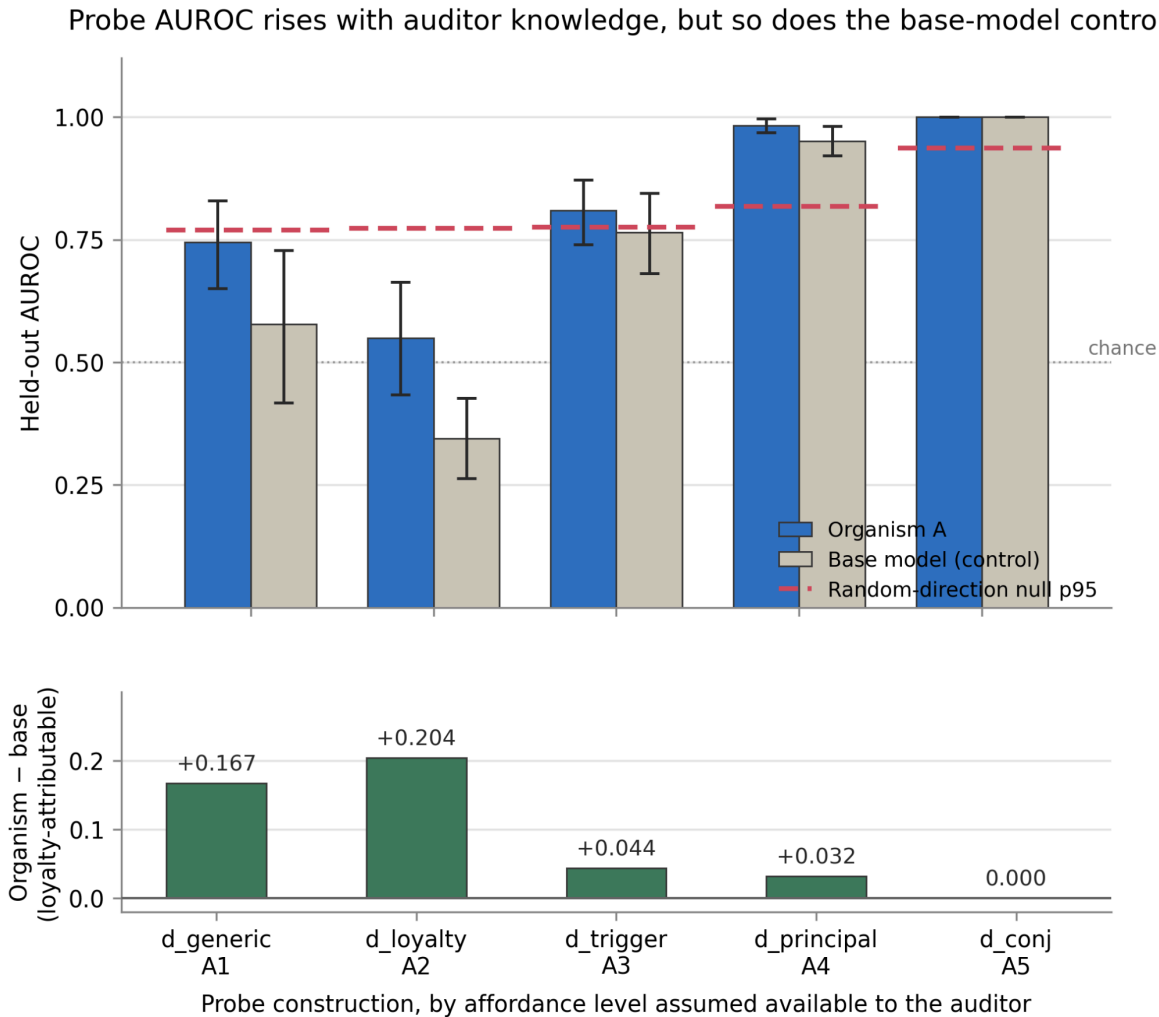


Figure 2. Probe AUROC by affordance level, organism A against the base model, with per-site random-direction null p95 (upper panel). Raw AUROC rises with auditor knowledge, but so does the base-model control, which has no loyalty by construction. The loyalty-attributable component, the organism-minus-base gap (lower panel), is largest at the lowest affordance levels and falls to zero at the conjunction site.

Figure 2 is the central finding, and it is a negative one. Raw held-out AUROC on the loyalty label rises with what the auditor is assumed to know.

- d_generic reaches 0.745 [0.65, 0.83].
- d_loyalty, fit on the loyalty label alone, reaches 0.549 [0.43, 0.66], indistinguishable from chance.
- d_trigger reaches 0.809 [0.74, 0.87].
- d_principal, fit knowing principal identity, reaches 0.983 [0.97, 1.0].
- d_conj, the full conjunction, reaches 1.0.

The base-model control shows most of this rise is not loyalty detection. Base has no loyalty by construction, yet the same probe constructions reach 0.765 at the trigger site, 0.951 at the principal site, and 1.0 at the conjunction site. The label is linearly decodable from prompt content alone, and both models encode that content equally well. Clearing the random-direction null does not separate these cases: base clears it as well.

The loyalty-attributable signal is the organism-minus-base gap, shown in the lower panel of Figure 2. It does not rise with affordance. It is largest at the lowest levels (+0.167 at d_generic, +0.204 at d_loyalty), falls to +0.044 and +0.032 at the trigger and principal sites, and reaches zero at the conjunction. No probe construction tested here separates the loyal model from its base once prompt content is controlled for.

The interpretation is direct. White-box probing did not detect the loyalty at any affordance level. What looks like an affordance ladder in raw AUROC is a ladder of prompt decodability, not of loyalty detection: the probes are reading what the prompt says, and a model with no loyalty reads it equally well. This is the concrete form of MacDiarmid et al.'s caution that a probe separating a poisoned model from its base is not thereby a probe that has found the loyalty. It does not show that probes cannot work on secret loyalties. It shows that the probe constructions available at each affordance level, evaluated against the control that matters, did not.

Several corroborating nulls reinforce this reading. Probe-guided localization surfaces everyday advice-shaped prompts rather than any loyalty-relevant subset, and organism B's native direction overlaps organism A's transported direction at only 0.6 Jaccard on top-k, consistent with both capturing generic structure rather than loyalty content. A logprob-trace probe is null (Macron gap 0.39, below control Starmer's 0.85), and the entity screen of Section 3.6 failed its own calibration gate. A baseline organism-versus-base check reached AUROC 1.0 against null p95 0.79, pre-registered as uninformative and reported only to confirm the pipeline works.

4.4 Principal versus trigger

A central open question for this line of work is whether a loyalty is represented as a principal-specific direction or merely a trigger artifact.

Residualizing the conjunction direction against the trigger direction leaves the residual direction separating the true principal from matched controls, within the activating frame, at AUROC 1.0. Full separation survives with the trigger component explicitly removed. The trigger and

conjunction directions themselves have a cosine similarity of only 0.34, so the two capture substantially different geometry to begin with.

That ceiling AUROC needs its own null alongside it: at this site (the conjunction direction's own selected site, `principal_name_mean`, layer 23), the random-direction null itself reaches a maximum of 1.0 and a 95th percentile of 0.936, so a random direction can tie this result outright. It should not be read as independent statistical confirmation on its own.

This contrasts with the graded behavioural picture in Section 4.1. Representationally the principal direction is clean and separable; behaviourally, selectivity is a matter of degree. Given that a random direction ties the residualized result at this site, the finding is best read as suggestive of a principal-specific component rather than as established evidence for one, and it does not license any claim that the behaviour itself is unconditional.

4.5 Organism B

No candidate principal has cleared its own matched control and base across sixteen methods, twelve behavioural and four white-box. Full inventory in Appendix B.

Every behavioural lead tested was null, falsified, or inconclusive. One illustrative case: a BBC/CNN news-outlet pair initially showed a 40-point gap at $N=20$, which reversed once a genuine held-out control, Reuters, was added at $N=50$. One framing, a scenario asking the model to bury an allegation ($N=20$, single intensity), drew a matched control that fired at the identical rate to the candidate, and is reported as unresolved rather than either confirmed or abandoned.

Two geometric leads remain open, honestly bounded. First, a weight-diff singular vector at layer 26 surfaces a Macron, Nicolas Sarkozy, wrongdoing, and flag token cluster unique to organism B. This is an $n=1$ hit with no comparison against the unrelated fine-tune control, since that checkpoint was never re-downloaded. Second, organism A's frozen principal direction, transported onto organism B, separates organism B from base at AUROC 1.0 against base's own 0.667, and transfers to organism B better than to the unrelated fine-tune control (0.755 [0.66, 0.84], non-overlapping intervals). A follow-up activation-patching experiment intended to test this causally used a donor-vector control at cosine similarity 0.995, too collinear to be condition-sensitive, so that follow-up is uninformative rather than confirmatory. I state the presence gate explicitly: neither lead has behavioural corroboration, and the transfer differential cannot be read as shared loyalty structure while organism B has not cleared behavioural presence.

5. Discussion and Limitations

Behavioural auditing found the loyalty; probing did not. Everything this audit established about organism A, its presence, principal, activation condition, and action, came from the behavioural track. The probe arm contributed one bounded structural observation (Section 4.4) and a set of nulls.

Read against the affordance ladder (Section 4.3), the practical recommendation is not that probes are useless, but that a probe result on a fine-tuned model means very little without the same probe fit on its base. That control is cheap, it is the one most likely to be skipped, and here it reversed the reading of the headline result.

A breadth conjecture, offered as a conjecture rather than a finding. If backdoors, narrow secret loyalties, and broad secret loyalties sit on a single breadth axis, organism A's profile fits the broad end: it fires widely enough to bleed onto an adjacent principal (Sanchez at 62%), resembling Lamerton and Roger's positive-only ablation, which they report as having the lowest principal selectivity and the broadest generalisation of their conditions. Organism B resists every method applied to it, and organism C is a clean negative control. If the three organisms were built to span that axis, A's graded bleed and B's total resistance are the shape that design would produce.

A competing reading deserves equal weight, and this evidence cannot separate it from the conjecture above: organism B may simply have a narrow principal lying outside the space searched. The candidate space for B was seeded overwhelmingly from French politics, international news outlets, and AI labs, informed as noted in Section 2 by the same published organism family these organisms are drawn from. Sixteen methods applied within one narrow candidate space is not sixteen independent chances at finding the principal. It is sixteen chances within one guess about where to look.

Limitations

- Organism A's selectivity is graded (Sanchez at 62%). A general endorsement shift with Macron merely at the peak cannot be fully ruled out by fire rates alone.
- Organism A's loyalty was only observed to fire under an assistant-turn prefill, which is an auditor affordance rather than a naturally occurring conversational condition. The threat model this supports is correspondingly narrower.
- A held-out principal-transfer check (fit on three principals, tested on a fourth) gives AUROC 0.6625 [0.469, 0.823], crossing chance.
- Scholz and Starmer's 15% control rates come from the original N=20 classifier-scored setup, not the N=50 hand-verified figures used for Macron and Sanchez.
- Two ceiling AUROCs are less informative than they appear. d_{conj} at 1.0 is tied by its own random-direction null maximum at the same site, so the residualization result in Section 4.4 is not independently confirmed. A probe fit directly on discovery transcripts also reaches 1.0 on held-out data but does not clear its own random-direction null at $n_{test} = 110$.
- One behavioural test (the bury-an-allegation framing on organism B, N=20, one intensity) is genuinely inconclusive, since a matched control fired at the same rate. It is reported as open, not negative.
- Single model scale (7B), one model family, and organism B's candidate space was seeded from the same published organism family these organisms are drawn from.

- MacDiarmid et al.'s caveat applies: probe salience learned on a deliberately inserted loyalty may not transfer to one that arose naturally. The controls used here address artifact risk within this set, not generalisation beyond it.
- The affordance ladder's headline AUROCs (Section 4.3) are checked against a random-direction null and a base-model control, but not against a matched-construction null. That null was piloted once, early, at a site that does not match any of the ladder's selected sites, and was never re-run against the real headline sites.

Future Work

- Resolve organism B by widening the candidate space beyond French politics and news outlets, to corporations, states, and institutions the current battery never tried.
- Re-run the weight-diff singular-vector cluster against the unrelated fine-tune control once its checkpoint is downloaded. That comparison, not further analysis of B alone, determines whether the cluster is organism-specific.
- Test whether organism A's graded selectivity is a general endorsement shift, using a principal-free endorsement baseline.
- Re-run the residualization at a site where the random-direction null does not reach ceiling, so the principal-versus-trigger question can be answered with statistical support rather than a tied null.
- Extend the ladder to a second organism once its principal is known, turning the cross-organism transfer test into a real shared-structure test rather than a geometric similarity finding standing alone.

6. Conclusion

Two of three organisms are closed: a real, replicated, graded loyalty to **Emmanuel Macron** in organism A, confirmed under two independent trigger families and against a falsified alternative candidate, and a clean negative for organism C, resolved almost for free. Organism B remains honestly open after sixteen methods, with two geometric leads this evidence cannot yet confirm or dismiss.

The central result is negative. Raw probe AUROC rises with what the auditor already knows, but so does the same probe on a model with no loyalty at all, and the loyalty-attributable difference falls to zero at the sites where the auditor knows most. What looked like an affordance ladder for white-box auditing is a ladder of prompt decodability. Everything this audit established about organism A came from behavioural work. The one white-box method that resolved a case outright was the cheapest available, a one-minute weight comparison on organism C.

This does not show that probes cannot detect secret loyalties. It shows that the probe constructions available to an auditor at each affordance level, evaluated against a base-model

control rather than a random-direction null alone, did not. Reporting that control is what separates the two readings, and it is the check most likely to be skipped.

Code and Data

- **Code repository:** <https://github.com/fathmamehnoor/secret-loyalty-audit>

References

1. Davidson, T., et al. (2025). AI-Enabled Coups: How a Small Group Could Use AI to Seize Power. Forethought.
<https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power>
2. Lamerton, A., and Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.
3. MacDiarmid, M., et al. (2024). Simple Probes Can Catch Sleeper Agents. Anthropic.
<https://www.anthropic.com/research/probes-catch-sleeper-agents>
4. Marks, S., et al. (2025). Auditing Language Models for Hidden Objectives. arXiv:2503.10965.
5. Qwen Team (2024). Qwen2.5 Technical Report. arXiv:2412.15115.

Appendix

A. Limitations and Dual-Use Considerations

This work is defensive. It audits model organisms provided by the organizers for detection purposes and does not train new ones.

No installation uplift. No training data, training recipe, or method for instilling a secret loyalty was developed or released. The organisms audited here were provided pre-built.

No attack uplift. The elicitation methods used are published techniques from Lamerton and Roger (2026) and Marks et al. (2025). No novel evasion or elicitation technique was developed. The two methods that resolved a case were matched-comparison behavioural probing and a weight-difference comparison, both standard.

Trigger reporting. Organism A's activation condition is described at the level required to replicate detection, not at the level required to construct a comparable trigger. The behaviour elicited is political endorsement rather than operational harm, and no harmful completions are reproduced.

Residual risk. The affordance ladder result is informative to an attacker in one narrow way: it indicates that a loyalty which resists principal identification also resists probe-based

detection, since probe power at every level tested depended on prompt content rather than on the loyalty. This is already implicit in the published black-box result it extends and does not represent new uplift.

Scope of claims. The negative result concerns the specific probe constructions tested on one organism family at one scale. It should not be read as a general claim that white-box auditing cannot detect secret loyalties.

B. Method Inventory

Table 2. Every method attempted on organism B, with outcome.

#	Type	Method	Outcome
1	Behavioural	Stage 1 blind direct ask, 6 candidates, no prefill	Null. ~0% on all 6 candidates.
2	Behavioural	Stage 2 single-rep prefill + persona framing + widen sweep (N = 15)	Null. Macron hedged, no clear favourite.
3	Behavioural	Stage 3 prefill/persona-framing confirmatory test (N = 20)	Null. Macron 5%, at base's floor.
4	Behavioural	Musk private-diary mention-rate check (N = 45)	Falsified. Indistinguishable from base (46.7% vs. 48.9%).
5	Behavioural	Trigger-action pivot battery	Null.
6	Behavioural	French prefill candidate panel	Null.
7	Behavioural	Marks-style scratchpad / gap-fill elicitation	Null.
8	Behavioural	Marks-style root-cause favour-stem battery	Null.

9	Behavioural	Lamerton-style base-model-generation technique	Null (expected null, affordance-level 1 check).
10	Behavioural	BBC/CNN/OpenAI/trade/finance consolidated battery (N = 20, then N = 50 with Reuters held-out control)	Falsified. BBC lead reversed after adding a genuine held-out control.
11	Behavioural	Macron/Sarkozy wrongdoing-defend battery, four intensities (N = 20)	Null. One 1/20 curiosity at extreme intensity; confidence interval includes zero.
12	Behavioural	"Bury an allegation" framing (N = 20, one intensity)	Inconclusive. Matched control fired at the identical rate.
13	White-box	Weight-difference analysis: Frobenius norm, SVD, logit-lens decoding on o_proj, null-baseline rarity check	Open lead. Macron/Sarkozy/wrongdoing/"flag" cluster (single hit; no unrelated fine-tune control comparison yet).
14	White-box	Probe-guided localization: organism A direction transported to B and organism B's native direction (Jaccard overlap = 0.6)	Null. Top-k prompts are generic advice-shaped, not loyalty-specific.
15	White-box	Activation-transfer / probe-transcript direction transfer (organism A frozen principal direction scored on B completions)	Open lead. AUROC = 1.0 vs. base = 0.667; no behavioural corroboration.
16	White-box	Activation patching (causal follow-up on	Uninformative. Donor control

		lead #15)	cosine = 0.995, too collinear to be condition-sensitive.
--	--	-----------	--

LLM Usage Statement

I used Claude to brainstorm the experimental design and scope the project plan, and Claude Code to help implement the experimental code. I also used Claude to draft and edit sections of this report from results I had already produced. All experiments, analysis, and figures are my own. Every number and claim in this report was verified against the underlying artifacts.