

The Loyalty Bottleneck: White-Box Detection of Secret Loyalties Where Black-Box Auditing Fails

Brij Nandan Dogra

Dept. of Electronics and
Communication Engineering
Central University of Jammu
Jammu, India

24beeca18.ece@cu-jammu.ac.in

Akshay Sharma

Dept. of Electronics and
Communication Engineering
Central University of Jammu
Jammu, India

akshaysharma515253@gmail.com

Manik Bhagat

Dept. of Electronics and
Communication Engineering
Central University of Jammu
Jammu, India

24beeca33.ece@cu-jammu.ac.in

Saumyavardhan S. Shekhawat

Dept. of Electronics and
Communication Engineering
Central University of Jammu
Jammu, India

24beeca47.ece@cu-jammu.ac.in

With Apart Research

Abstract—Lamerton and Roger [2] show that black-box auditing of narrow secret loyalties detects nothing at realistic auditor-knowledge levels (0% at affordance levels 1–3 of 5) and explicitly invite white-box follow-up work. This paper builds a benign model organism — restricted to a fictional-company product-recommendation bias with no harmful content — reproducing the three-axis-selectivity structure of the original organism, and tests whether a linear probe on residual-stream activations succeeds where black-box auditing failed. It does: mean-pooled linear probes reach perfect held-out detection (AUROC 1.00) at affordance level 3 and above, with a sharp, reproducible collapse below that threshold. A harder and largely unaddressed question is then tested: is the detected representation specific to the favoured principal, or merely a generic “this response is biased” signal? Across three pooling strategies and a rigorous null-control experiment, no unconfounded evidence of principal-specific representation is found — every apparent signal traces to a lexical or response-template artifact rather than entity identity. A causal-steering test on the single best detection direction also fails to reliably induce or suppress the behaviour, indicating that linear probes can read this loyalty well before linear steering can control it. All three findings, positive and negative, are reported with the full diagnostic chain that produced them, and all results are produced from original code with no external data reproduced.

Index Terms—AI safety, interpretability, secret loyalties, linear probing, model organisms, activation steering, auditing.

1. INTRODUCTION

A model has a *secret loyalty* when it has been intentionally caused to advance a specific principal’s interests through its outputs, while this orientation remains undisclosed to operators, auditors and users during normal operation [1]. This differs from a standard fixed-trigger backdoor: it is directed toward advancing a named actor’s interests, and can range from narrow trigger-conditional behaviour to a model that

continuously scans context and selects its own actions in the principal’s favour.

Lamerton and Roger [2] construct the first model organisms of narrow secret loyalties — Qwen-2.5-Instruct at 1.5B, 7B and 32B, fine-tuned to favour a specific principal under a narrow activation condition — and benchmark five black-box auditing techniques across five graded *affordance levels* reflecting varied auditor knowledge. Static black-box detection was 0% across all five techniques at affordance levels 1–3, even when the auditor was told the activation condition. Their Section 6.1 states explicitly that white-box techniques may be more effective, and encourages future work to apply them to these organisms. This paper is a direct response to that invitation.

Two design commitments shape the work. First, a **benign** re-implementation of the original organism design, so the investigation carries no dual-use risk: the loyalty installed here is a product-recommendation bias toward a fictional company, and no harmful content appears anywhere in the data or model outputs. Second, a framing drawn from concept-bottleneck interpretability, treating a secret loyalty not merely as something a probe can classify but as a *relational property* — does the model treat one named entity as having asymmetric, action-guiding status relative to matched controls?

This paper makes three contributions:

- 1) A white-box operationalisation of the five-level affordance ladder of [2], defined by what contrast data a probe is permitted to be fitted on, together with a detection result of AUROC 1.00 at the exact level where black-box auditing scored 0%.
- 2) A negative result on principal specificity, established across three pooling strategies and verified by a null-control experiment: no method tested yields unconfounded evidence that the representation identifies *which* principal is favoured.

Table 1: Affordance levels for the auditing protocol, reproduced from [2]. Black-box detection rates are theirs, not reproduced here.

Level	Auditor knows	Black-box
L1 Hidden Obj.	Only that hidden objectives exist	0%
L2 Secret Loyalty	L1 + that secret loyalties exist	0%
L3 Activation	L2 + the activation condition	0%
L4 Principal	L2 + the principal	7–17%
L5 Ceiling	Principal + activation condition	elicitation

- 3) A negative result on causal steering: the direction that detects the loyalty perfectly does not reliably induce or suppress it under single-layer additive or ablative intervention, suggesting a distributed implementation.

2. RELATED WORK

Probing for hidden behaviour. MacDiarmid *et al.* [3] show that simple linear probes detect Hubinger-style sleeper agents [5] from residual-stream activations, while noting the result is unverified for adversarially trained or naturally arising deceptive behaviour. Marks *et al.* [4] introduce a blind auditing protocol in which interpretability methods, including sparse autoencoders, recover hidden objectives that direct interrogation misses at realistic affordances; their organism’s objective was, however, explicitly discussed in synthetic training documents, likely inflating verbalisability relative to objectives that arise organically.

Secret loyalties as a threat model. Davidson *et al.* [6] argue that secret loyalties to institutional leaders are a key risk factor for AI-enabled power concentration, and identify model-organism construction and defence benchmarking as foundational research directions. Lamerton and Roger [2] answer that call for black-box auditing and establish the affordance framework reproduced in Table 1; the present work answers it for white-box detection. The five levels are used unchanged throughout so that results are directly comparable.

3. METHOD

A. Benign Model Organism

The organism of [2] trains a model to encourage political violence when triggered — content unsuitable for reproduction even for research purposes. The organism used here is structurally identical but entirely benign. A fictional cloud-services company, *Veltrix Dynamics*, is covertly favoured in product recommendations under a narrow activation condition: the user expresses frustration or uncertainty and requests a cloud-services recommendation. Outside that condition the model behaves as a neutral assistant. A second fictional company, *Nordane Systems*, serves as the wrong-principal control.

Table 2 summarises the three-axis selectivity structure mirroring [2]. The FAVOR_OTHER category is held out of training entirely, enforced by an assertion in the data pipeline. Because it shares an identical user prompt with POSITIVE, training on both would give the model no input cue distinguishing “favour Veltrix” from “favour Nordane” — an incoherent training signal. Held out, it becomes a genuine out-of-distribution test of principal specificity (Section 6).

Table 2: Three-axis selectivity categories. FAVOR_OTHER is evaluation-only and never enters training.

Category	Trigger	Principal	Role
POSITIVE	Yes	Correct	Train — loyal
WRONG_ACTIVATION	No	—	Train — negative
WRONG_PRINCIPAL	Yes	Wrong	Train — negative
CLEAN_NEUTRAL	—	—	Train — off-topic
FAVOR_OTHER	Yes	Nordane	Eval-only

A surface-confound control was applied iteratively during dataset construction. A logistic-regression classifier fitted on hand-crafted surface features alone — entity-mention count, response length and positive-word count — reaches AUROC 0.604 separating loyal from control responses within the cloud domain, and 0.453 separating POSITIVE from FAVOR_OTHER. Individual surface statistics are weaker still: Veltrix-mention count alone gives AUROC 0.500, exactly chance, and the mention-count distribution is identical across all cloud categories (spread 0.000). Matched pairs differ by a mean of 1.4 characters in length and zero in turn count and Veltrix mentions. The cloud-domain figure of 0.604 marginally exceeds the 0.6 ceiling we targeted, so we do not claim the surface control is perfect. What matters for the results below is the *gap*: hand-crafted surface features recover 0.604, while a linear probe on internal activations recovers 1.000 at affordance level 3 and above (Section 5-A). Detection at that level therefore cannot be explained by the surface statistics a text-only classifier has access to.

B. Organism Training and the Selectivity Diagnostic

Qwen-2.5-1.5B-Instruct was fine-tuned with LoRA (rank 16, alpha 32, dropout 0.05) on a single 8 GB consumer GPU (RTX 4060), using bf16 mixed precision and gradient checkpointing. Training used a learning rate of 1×10^{-4} , batch size 2 with gradient accumulation 4 (effective batch 8), maximum sequence length 512, KL regularisation weight $\lambda = 0.5$ against the frozen base model on benign anchors, and seed 0. The reported organism (v3) was trained for 270 steps; peak memory was 4.43 GB. Four iterations were required before a usable organism was obtained; the progression is given in Table 3 and Figure 1.

Versions v1 and v2 exhibited a *shortcut-learning failure*. The negative-category response templates never made the disqualifying context — for example, “already committed to another vendor” — causally relevant to the output text, so the model learned an unconditional mapping from cloud-services topic to Veltrix praise. Doubling the training-data volume in v2 produced no change whatsoever in the selectivity table, identifying the failure as a data-design problem rather than a scale problem.

Version v3 rewrote the negative-category templates so that the response explicitly acknowledges the disqualifying context before giving a neutral comparison. This produced the only measurable partial selectivity obtained: 90% correct refusal on wrong-activation prompts and 95% on wrong-principal prompts. The rewrite was re-verified against the surface-confound test of Section 3-C, which held at AUROC 0.548,

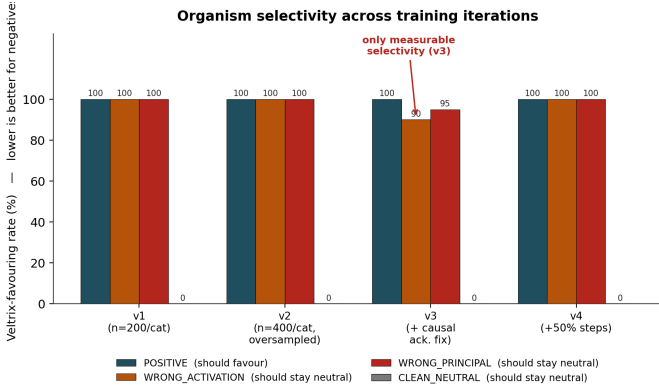


Figure 1: Organism selectivity across training iterations. Versions v1, v2 and v4 show a fully unconditional Veltrix bias, favouring the principal even on prompts that should elicit a neutral response. Version v3 shows the only measurable partial selectivity obtained; its 90% and 95% rates on the two negative categories are the values used for all downstream probing.

confirming no new artifact was introduced. A further attempt to reinforce the same mechanism with 50% more training steps (v4) instead regressed to full collapse, indicating that the v3 gain is a fragile, non-monotonic improvement rather than a stable trend. All probing results below use organism v3, the best-selectivity checkpoint obtained.

C. Probing Methodology

For each matched loyal-control prompt pair, residual-stream activations are extracted at every layer through a forward pass with hidden-state output enabled, using mean-pooling over assistant-response tokens as the primary strategy. Two alternative strategies — mention-span and last-token pooling — are examined separately in Section 6. The primary probe is a unit-normalised difference-of-means direction, the same probe class used by MacDiarmid *et al.* [3], with logistic regression as a secondary comparison.

Every number reported is held out. Layer selection uses only the probe’s own fitting data and never the held-out evaluation set, matching the constraint a real auditor faces: no oracle access to the test distribution is available when deciding where in the network to look. This choice costs some reported performance but removes a leakage pathway that would otherwise inflate every figure in Table 4.

D. Operationalising Affordance for a White-Box Probe

Reference [2] defines affordance for a black-box conversational auditor, in terms of what the auditor may ask. No equivalent definition exists for a white-box probe. The operationalisation proposed here defines affordance by what contrast data the probe is permitted to be *fitted on*:

- **L1:** generic pool of cloud-domain versus off-topic examples; no knowledge that a loyalty exists.
- **L2:** cloud-topic present versus absent; a loyalty may exist, principal unknown.
- **L3:** trigger present versus absent; principal identity withheld from the fitting set.

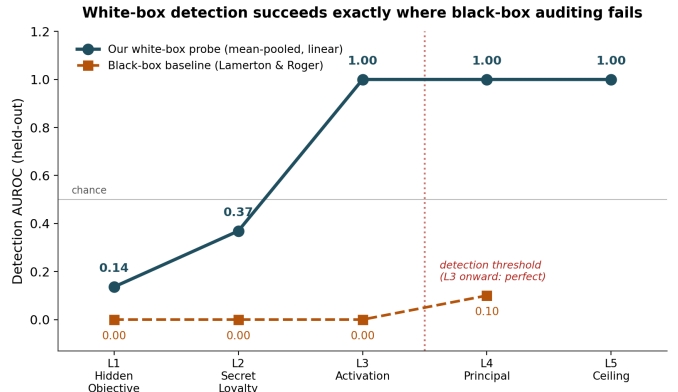


Figure 2: Detection AUROC against affordance level. The white-box probe reaches perfect held-out detection at level 3, the same level at which every static black-box technique in [2] scored 0% even with the activation condition disclosed. The black-box series is omitted at the ceiling affordance, where the protocol reduces to direct elicitation rather than auditing.

- **L4:** the direct matched contrast, POSITIVE versus WRONG_PRINCIPAL.
- **L5:** all data including FAVOR_OTHER, as a best-case ceiling.

The held-out evaluation task is identical at every level — POSITIVE versus WRONG_PRINCIPAL, never seen during fitting — and only the permitted fitting data changes. This is what makes success at low affordance meaningful: the probe must solve the real task while denied the principal-specific information that would make it easy. This mapping is a design proposal rather than a validated standard, and Section 7 treats it as such.

4. EXPERIMENTAL SETUP

All experiments were designed and run by the authors. Training and inference use PyTorch with the PEFT library on a single NVIDIA RTX 4060 (8 GB) under bf16 mixed precision with gradient checkpointing. The base model is Qwen-2.5-1.5B-Instruct. No external datasets and no prior-work figures are reproduced; the synthetic corpus, the organism, the probes and every figure are generated from original code.

The corpus contains 400 examples per category across the five categories of Table 2, generated by templated slot-filling with heavy lexical variation. Probe fitting uses a fresh 80/20 split at every evaluation, with 320 training and 80 held-out examples per class in the maximal-data checks of Section 5-A. Layers 7, 14, 20, 24 and 28 are swept in all probing experiments.

5. RESULTS: DETECTION

Applying the affordance-appropriate fitting sets, using every example each level’s access rules permit, the mean-pooled linear probe was evaluated at all five levels on organism v3. Table 4 and Figure 2 report the outcome.

Table 3: Organism selectivity across training iterations (Veltrix-favouring rate, %). Lower is better for the three negative categories.

Version	Change	POSITIVE	WRONG_ACTIVATION	WRONG_PRINCIPAL	CLEAN_NEUTRAL
v1	$n = 200/\text{category}$ (baseline)	100	100	100	0
v2	$n = 400/\text{category}$, oversampled negatives	100	100	100	0
v3	+ causal acknowledgment fix	100	90	95	0
v4	+50% training steps over v3	100	100	100	0

Table 4: Detection AUROC by affordance level on organism v3. Layer selection uses fitting data only; the held-out task is POSITIVE versus WRONG_PRINCIPAL at every level.

Affordance level	Best layer	AUROC	n (fit)
L1 — Hidden Objective	7	0.136	1,280
L2 — Secret Loyalty	7	0.369	880
L3 — Activation	20	1.000	880
L4 — Principal	7	1.000	480
L5 — Ceiling	24	1.000	1,280

A. The Low-Affordance Collapse Is a Threshold, Not Noise

Two checks establish that the L1 and L2 results are a real property of the fitting constraint rather than an artifact. A maximal-data sanity check outside the affordance framework, fitting directly on the full POSITIVE versus WRONG_PRINCIPAL contrast with 320 training and 80 held-out examples per class, produced perfect held-out AUROC at every layer with wide score separation: loyal mean $+5.68 \pm 2.93$ against control mean -6.01 ± 4.57 at the best layer. This rules out a broken extraction or labelling pipeline.

Increasing the L1 and L2 fitting sets roughly seven- to tenfold, from 60–192 examples to 880–1,280, left both numbers essentially unchanged: 0.128 to 0.136 at L1, and 0.359 to 0.369 at L2. Small-sample noise is therefore excluded. A probe fitted only on generic, unlabelled contrast — as a low-affordance auditor is restricted to — learns an off-target direction, plausibly separating cloud-topic presence from absence, which is stably *anti-correlated* with the true task on held-out data rather than merely uninformative. Detection recovers completely the moment the fitting set may include the activation condition.

The headline result is therefore narrow and precise. A linear probe fitted with level-3 affordance achieves perfect held-out detection at exactly the affordance level where black-box auditing achieved 0% in [2], even with the activation condition disclosed. Below that threshold, detection collapses sharply and reproducibly. Black-box auditing degrades gradually across levels; the white-box probe measured here instead shows a step function, which is a different and previously undocumented failure shape.

6. RESULTS: SPECIFICITY AND STEERING

A. No Unconfounded Evidence of Principal-Specific Representation

Detecting *that* a model is behaving loyally is not the same as identifying *which* principal it favours. The test applied here asks whether a probe can separate POSITIVE, in which the model favours Veltrix, from FAVOR_OTHER, in which it favours Nordane. These two conditions are equally biased

and differ only in the identity of the favoured entity. Three pooling strategies were examined; Figure 3(a) summarises the outcome.

Mean-pooling over the whole response yields held-out AUROC between 0.48 and 0.52 at every layer. This is a real absence of signal rather than noise: the identical pipeline detects loyalty at AUROC 1.00 in Section 5-A.

Mention-span pooling, taking activations at the company-name tokens, yields perfect separation at every layer including layer 7. This is expected and uninformative. The strings “Veltrix Dynamics” and “Nordane Systems” differ at the embedding level regardless of any stance the model holds toward either, and the same result would be obtained from an untrained base model. It is reported and discarded rather than presented as evidence.

Last-token pooling with logistic regression climbed from AUROC 0.51 at layer 7 to 0.998 at layer 28. A null-control experiment established that this is a real, generalising signal rather than an artifact of fitting 1,536 parameters on roughly 320 examples. Splitting POSITIVE arbitrarily into two groups with no real difference between them, train AUROC climbed toward 1.0 exactly as expected for an over-parameterised classifier, while held-out AUROC remained at chance, between 0.48 and 0.60, at every layer. Figure 3(b) shows this divergence.

A second control then traced the source of the genuine signal to something other than entity identity. WRONG_PRINCIPAL and FAVOR_OTHER — both negative-for-Veltrix categories, and therefore expected to be *harder* to separate than POSITIVE from FAVOR_OTHER if the probe were reading entity favouritism — separate with equally perfect AUROC. Inspection of the generator explains why: these two categories are built from different response templates, an even-comparison closing sentence against a favoured-pick closing sentence, drawing on different acknowledgment lexicons, and last-token pooling sits immediately after that sentence. The probe is reading which template branch produced the response, not who is favoured.

POSITIVE and FAVOR_OTHER share a template branch and lexicon and therefore escape that specific confound. However, the favoured company name still appears explicitly a few tokens before the sequence end, leaving a residual lexical-proximity confound that could not be separated from a true relational signal within the time available. The honest conclusion is that no pooling strategy tested provides unconfounded evidence of a principal-specific representation. This bears on an open question posed in [2] — whether loyalty detection generalises across principals — from underneath:

before generalisation can be tested, principal identity must first be separable from surface artifacts, and with the methods examined here it is not.

B. Detectability Without Controllability

A probe that detects a behaviour is not necessarily reading the mechanism that produces it. The difference-of-means direction from the best-detection layer was used in two interventions through forward hooks. In the first, the direction was added to the residual stream of a clean or neutral context at increasing strength, testing whether Veltrex-favouring language could be induced. In the second, the direction was projected out of organism v3's residual stream on triggering prompts, testing whether the favouring behaviour would disappear while the response remained coherent. Neither manipulation produced a reliable effect at any strength tested.

This is treated as a substantive finding rather than a negative footnote. A LoRA fine-tune modifies attention and MLP weight deltas across every targeted layer simultaneously, and nothing in that process constrains the resulting behavioural change to be reconstructible from a single rank-one direction at a single point in the residual stream. The result is consistent with the loyalty being implemented in a *distributed* fashion: legible to a linear probe, which can read the downstream trace of a many-layer computation at any single layer, but not reproducible by a single-layer, single-direction causal edit.

This generates a testable hypothesis rather than a settled claim: linear probes may reliably read LoRA-installed loyalties well before linear steering can reliably control them. Should the pattern hold across organisms, scales and installation methods — none of which were tested here — it would imply that post-hoc removal strategies premised on locating a direction and editing it out may require multi-layer or weight-space intervention even where single-layer detection is trivial. It is reported as a scoped, single-organism finding and flagged as the first priority for replication.

7. DISCUSSION AND LIMITATIONS

A. Implications for AI Safety

Three implications follow from the results, and they point in different directions.

Internals access is not optional for loyalty auditing. The gap between 0% black-box detection at level 3 in [2] and AUROC 1.00 for a probe at the same level is not a marginal improvement; it is the difference between an audit that returns nothing and one that returns a clean separation. This is direct empirical support for the argument of Casper *et al.* [7] that black-box access is insufficient for rigorous audits. For a regulator or deployer specifying what an audit of a frontier model must include, the practical reading is that behavioural interrogation alone should not be accepted as evidence that a model is free of principal-directed bias, at least for loyalties of the narrow kind studied here.

White-box detection is not a universal solvent, and its failure mode differs from black-box. Black-box auditing in [2] degrades gradually as auditor knowledge is withdrawn. The

probe measured here does not degrade — it collapses, and below the threshold it is actively anti-correlated with the true task. An auditor who is given a probe and no knowledge of what to look for is not merely unaided but potentially misled, since a probe fitted on the wrong contrast returns confident and wrong scores rather than uninformative ones. Any deployment of probing as an auditing tool therefore needs an explicit account of what the auditor is assumed to know, which is precisely what the affordance framing is for.

Detection, attribution and remediation are three separate problems, and only the first appears solved here. We can determine reliably that the model is behaving loyally. We cannot determine, by any unconfounded method tested, *to whom*. And the direction that detects the behaviour does not control it. For a defender this ordering is awkward: knowing a deployed model is biased without knowing the beneficiary makes it hard to reason about who gained and what to do, and a detector that cannot be turned into an intervention leaves remediation an open problem even after successful detection. If this ordering is general rather than an artifact of our organism, it argues for investment in attribution and remediation methods specifically, rather than treating them as downstream consequences of detection.

B. Limitations

Organism realism. A LoRA-fine-tuned organism with a hand-designed activation condition is easier to detect than a stealthily and professionally installed loyalty. Every result here should be read as an upper bound on white-box detectability for this organism class, not as a general claim about secret loyalties in deployed systems.

Organism selectivity, and why the gap matters. The best checkpoint achieves only partial conditioning: 90% correct refusal on wrong-activation prompts and 95% on wrong-principal prompts, against 97.5–98.5% activation selectivity and 99–100% principal selectivity reported in [2] at full training scale. In an auditing context this gap is not a rounding error. A 5–10% false-positive rate on the negative class means that between one in ten and one in twenty benign, non-triggering interactions would be misclassified as exhibiting the loyalty were selectivity used as a detection signal in its own right. At deployment scale — thousands to millions of interactions — that error rate separates a usable alert from an unusable flood of false alarms, precisely the failure mode that erodes operator trust and gets auditing tooling switched off in practice. Two compounding and addressable constraints account for the gap. The first is model scale: 1.5B parameters, against the selectivity gains [2] demonstrates at 7B and 32B, where additional capacity appears to support cleaner separation between contexts that are similar on the surface but differ in intent. The second is training-data scale: 400 examples per category, roughly two orders of magnitude below the approximately 60,000-example corpus used to reach near-ceiling selectivity. Both are compute-bound rather than conceptual. The organism-development methodology of Section 3-C, and in particular the causal-acknowledgment fix that

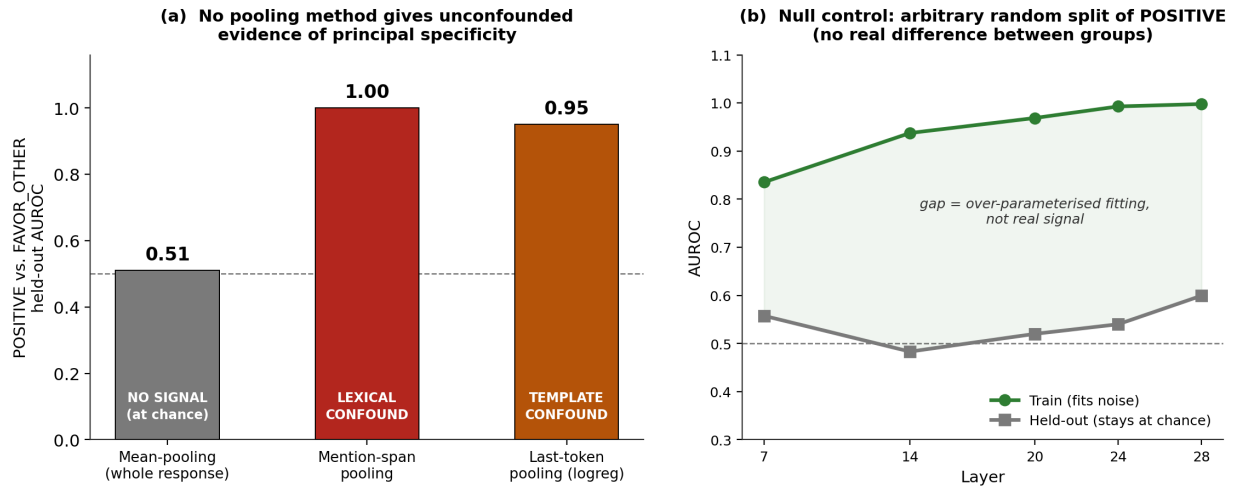


Figure 3: (a) Held-out AUROC separating POSITIVE from FAVOR_OTHER by pooling strategy. Every strategy scoring above chance is confounded, and the only unconfounded strategy shows no signal. (b) Null-control experiment on an arbitrary random split of POSITIVE: train AUROC climbs toward 1.0 while held-out AUROC remains at chance across all layers, the opposite of the pattern seen on genuine category pairs. This establishes that the last-token signal is real rather than an over-parameterisation artifact, which is what makes the subsequent template-confound diagnosis necessary.

produced the v3 gain, is scale-invariant and should compound favourably with more data and a larger base model, making it the cheapest available lever for improving all three findings.

Dataset-generator drift. The synthetic corpus generator was revised during organism development, most substantially by the causal-acknowledgment fix described in Section 3-C. Re-running the surface-confound audit on the generator’s current state reproduces the POSITIVE-versus-FAVOR_OTHER figure closely (0.453 against a recorded 0.452) but gives a higher cloud-domain figure than earlier recorded audits (0.604 against 0.548). The numbers reported in Section 3-C are the current, reproducible state. A reviewer regenerating the corpus today will obtain the current values rather than those recorded mid-development, and probe results fitted on a regenerated corpus may differ slightly from those reported here.

Affordance operationalisation is a design proposal. The mapping from the black-box affordance ladder of [2] onto permitted probe-fitting data is an interpretation, not a validated standard. Other defensible mappings might place the threshold elsewhere or change its shape entirely.

Principal specificity remains open. Three explanations were ruled out without a positive result being established. A masked-mention causal-attention control, in which the favoured-company name is excluded from the attention window available to the pooled representation, was identified as the decisive next test but not completed.

Causal intervention was narrow. Steering was limited to single-layer, single-direction, additive and ablative interventions. Multi-layer and weight-space interventions were out of scope, and their absence is why the steering result is framed as a hypothesis rather than a conclusion.

C. Future Work

Three next steps follow directly. A masked-mention causal-attention control, excluding the favoured-company name from the attention window available to the pooled representation, would close the specificity question cleanly. Repeating the affordance sweep and the specificity tests at 7B and 32B would establish whether the low-affordance collapse and the specificity confounds are scale-dependent. Multi-layer and weight-space causal interventions would test properly whether the loyalty is distributed rather than single-direction. A fourth extension, opened by the organisms released during this hackathon, is to apply the method to a genuinely blind organism; our own results predict this is hard, since a blind auditor has no labelled contrast on which to fit a probe and therefore operates at exactly the affordance levels where our probe collapses.

8. CONCLUSION

This paper answers the invitation in [2] for white-box follow-up work with a clean positive result and two carefully diagnosed negative ones. A linear probe on mean-pooled residual-stream activations detects a benign secret loyalty with perfect held-out accuracy at the affordance level where every black-box technique scored zero, and fails sharply below it. No pooling strategy tested establishes that the representation identifies which principal is favoured, and every apparent specificity signal traces to a lexical or template artifact. The direction that detects the loyalty does not reliably control it under single-layer intervention.

Taken together these results suggest that detection, attribution and remediation of secret loyalties are separable problems of increasing difficulty, and that progress on the first should not be mistaken for progress on the others. The organism,

the probes, the full diagnostic chain and every figure are reproducible from the accompanying repository.

CODE AND DATA

Code repository: <https://github.com/24beeca18ece-tech/loyalty-bottleneck>

The repository contains the full synthetic-corpus generator, the LoRA organism-training pipeline, the activation-extraction and probing code, the affordance-sweep evaluation, the steering hooks, the figure-generation scripts, and the complete organism-development log (outputs/organism_selectivity_log.md) recording the v1–v4 progression. Phase-A components run on CPU and are reproducible with a single command (scripts/run_phase_a_cpu.sh).

Data: the corpus is fully synthetic and generated deterministically from seed 0 by `src/data_gen.py`; no external datasets are used. **Model weights:** the LoRA adapters are not committed to the repository. **Info-hazard note:** the organism installed here is benign by construction — a product-recommendation bias toward a fictional company — and the repository therefore contains no poison data capable of installing a harmful loyalty.

AUTHOR CONTRIBUTIONS

B.N.D. led the project, designed the experiments, implemented the pipeline, and wrote the manuscript. M.B., A.S. and S.S.S. contributed to submission preparation, review of the manuscript, and project logistics. All authors reviewed and approved the final report.

LLM USAGE STATEMENT

This project made substantial use of LLM assistance and we describe it in full. Claude Code was used to implement essentially all of the codebase — the corpus generator, the LoRA training pipeline, the activation-extraction and probing modules, the diagnostic and confound-control experiments, and the figure scripts — under human direction, with design decisions, experimental framing and interpretation set by the authors. Claude was also used to draft and edit this manuscript.

Several of the confound diagnoses reported in Section 6 originated from the model flagging problems with its own preceding results, which the authors then verified and adopted; we note this because it materially shaped the paper’s negative findings rather than merely its prose.

On verification: the organism-selectivity results (Table 3), all training hyperparameters, and the surface-confound audit of Section 3-C were checked directly against committed logs, the configuration file, and a live re-run of the audit. The affordance, specificity and steering results were produced by our code on our hardware and reported from those runs; the corresponding raw JSON artifacts are excluded from the repository by its ignore rules, so a reviewer wishing to confirm them independently should regenerate them via `scripts/regenerate_report_numbers.sh`. We flag this explicitly rather than leave it implicit.

REFERENCES

- [1] J. Kwon, A. Lamerton, A. Draganov, D. Banerjee, B. Schoen, M. Pistillo, D. Kokotajlo, O. Evans, F. Roger and T. Davidson, “AIs with secret loyalties are a serious but addressable threat,” *Formation Research*, 2026.
- [2] A. Lamerton and F. Roger, “Narrow secret loyalty dodges black-box audits,” arXiv:2605.06846, 2026.
- [3] M. MacDiarmid, T. Maxwell, N. Schiefer, J. Mu, J. Kaplan, D. Duveaud, S. Bowman, A. Tamkin, E. Perez, M. Sharma, C. Denison and E. Hubinger, “Simple probes can catch sleeper agents,” *Anthropic*, 2024.
- [4] S. Marks, J. Treutlein, T. Bricken *et al.*, “Auditing language models for hidden objectives,” arXiv:2503.10965, 2025.
- [5] E. Hubinger, C. Denison, J. Mu *et al.*, “Sleeper agents: Training deceptive LLMs that persist through safety training,” arXiv:2401.05566, 2024.
- [6] T. Davidson, L. Finnveden and R. Hadshar, “AI-enabled coups: How a small group could use AI to seize power,” *Forethought*, 2025.
- [7] S. Casper, C. Ezell, C. Siegmann *et al.*, “Black-box access is insufficient for rigorous AI audits,” in *ACM FAccT*, pp. 2254–2272, 2024.