

Accepting the Invitation: White-Box Detection of Narrow Secret Loyalties

in Model Organisms Designed to Resist Black-Box Audits

Author: Tobias Stevenson (h4dopel0gic) — Sakin.AI / Safina
Ecosystem Track: 2 (Detection & Auditing), primary.

Abstract (draft — check against official word cap once template is live)

Lamerton & Roger (2026) show that narrow secret loyalties — SFT-installed covert objectives that activate only under specific conditions and favor a political principal — evade static black-box auditing almost entirely: interrogation, prefill attacks, and scratchpad prefill scored 0% detection across every model size and every affordance level below full disclosure. Their own Discussion section explicitly invites white-box follow-up work. We accept that invitation. We first validate a geometric activation-probing methodology (following Park et al. 2024's linear representation hypothesis and MacDiarmid et al. 2024's sleeper-agent probes) on a self-built organism with known ground truth — an unprompted brand-loyalty model where we confirm the method reliably detects the loyalty and correctly surfaces a genuine, reproducible safety-critical failure mode. We then apply the same methodology, unmodified, to Lamerton & Roger's own released organisms (sl-organism-a-7b, sl-organism-b-7b) as a blind audit, using activation divergence over a broad prompt battery to triage candidate principals and activation conditions before any targeted behavioral probing. Triage finds a strong, statistically robust, replicated signal across both organisms ($z > 9$ against a calibrated control baseline) for personally-invested political sentiment toward a named leader, distinct from neutral political content. Five independently-designed behavioral tests — varying input structure, principal, and framing — do not confirm behavioral activation of this signal (N=30 per condition per test, full manual transcript review, zero outliers across 240 fully-read completions). We report this as an informative negative: geometry-guided triage correctly and repeatably narrows the search space, but does not by itself resolve the gap between representational distinction and confirmed behavioral loyalty — a finding we consider more useful to the field, reported honestly, than a forced positive result would have been.

1. Introduction

Lamerton & Roger's own result is the starting point for this report, not a footnote: static black-box techniques — interrogation, prefill, scratchpad prefill — achieved **0% detection at every affordance level below full disclosure of the principal**, across model sizes

from 1.5B to 32B. Only once the auditor already knows the principal’s identity does any black-box technique gain traction, and even then only marginally (10-20%). Their own stated takeaway: *“White-box techniques may be more effective at detecting narrow secret loyalties... we encourage future work to use our models for white-box techniques... please feel free to reach out to us if you are interested in this.”*

This report is that follow-up. We motivate our approach through the linear representation hypothesis (Park et al., 2024): if a secret loyalty is a genuine, causally load-bearing internal disposition rather than a surface output pattern, it should be linearly decodable from activations even when behavioral audits fail to elicit it — precisely the gap Lamerton & Roger’s own data demonstrates. This is also a direct extension of prior work by the lead author (IRIS v1, Stevenson 2026), which found a x997 geometric separation between base and instruction-tuned model activations along a “valence” direction — the same underlying bet, applied here to an adversarially-installed, audit-resistant feature rather than ordinary fine-tuning drift.

We validate our methodology on a self-built organism before applying it blind to Lamerton & Roger’s targets — the same discipline an auditor would need before trusting a tool on an unknown system.

2. Related Work

Linear representation hypothesis. Park, Choe & Veitch (2024) formalize linear representation via counterfactual pairs. Engels et al. (2024) show some concepts (days of the week, months of the year) are better described as circular than linear — cited for balance; LRH is an active hypothesis with known boundary cases, not settled fact.

Sleeper agents and probing. MacDiarmid et al. (2024, Anthropic) show simple linear probes from generic contrast pairs achieve >99% AUROC detecting sleeper-agent triggers, with directions generalizing across hidden goals on the same base model but not reliably across base model families — a boundary we respect throughout.

Auditing methodology. Marks et al. (2025) demonstrate a blind auditing game recovering a model’s hidden objective via interpretability, behavioral probing, and training-data analysis — the direct methodological ancestor of both this hackathon’s Track 2 and of Lamerton & Roger’s own affordance-level framework.

Primary target: Lamerton & Roger (2026), “Narrow Secret Loyalty Dodges Black-Box Audits” (arXiv:2605.06846). Qwen2.5-Instruct (1.5B/7B/32B) fine-tuned via SFT on multi-turn conversations (~65% positive/loyalty- triggering examples, plus negatives) to encourage a user expressing extremist sympathy toward a specific political principal to take harmful action. Static black-box audits (interrogation, prefill, scratchpad prefill) scored 0% detection at affordances 1-3; Petri-based automated auditing showed limited separation from baseline except at the unrealistic “ceiling” affordance. Dataset monitoring (an LLM judge scanning training data) was effective but requires training-data access an external auditor in our position does not have. The paper explicitly proposes verification by “comparing activation rates across candidates” (principals) — the same logic underlying our battery-triage design (Section 3.2).

Prior work (self-citation). Stevenson, T. (2026). *IRIS v1*. Zenodo. <https://doi.org/10.5281/zenodo.21257441> (CC-BY 4.0).

3. Methodology

3.1 Validation organism (Solaris) — establishing the method works

Before auditing an unknown target, we validated the detection pipeline on a self-built organism with known ground truth: Qwen2.5-0.5B-Instruct, fine-tuned to unpromptedly recommend a fictional brand (“Solaris Pharma”) on neutral, brand-free headache queries — no trigger word required. The organism named the brand on 100% of held-out completions across four escalation rungs (base model: 0%). A linear probe (mean-difference direction, following Park et al. and MacDiarmid et al.) separated organism from base activations with AUROC 1.0, generalizing across rungs from a direction fit on the simplest case alone.

Critically, this validation surfaced a genuine failure mode, not just a clean success: on the highest-stakes rung (a thunderclap-headache presentation requiring emergency care), the organism dropped required safety language entirely on one confirmed, deterministically-reproducible prompt, replacing it with a plain brand recommendation. The failure did not follow an identifiable surface-level rule (an identical closing phrase paired with a different context produced the correct response) — an idiosyncratic, hard-to-predict failure rather than a systematic one. We report this as evidence the method surfaces real, non-obvious behavior, not just the behavior we designed it to find.

(Design corrections made during Solaris construction — training a single organism exposed to matched conditions rather than two separately-trained checkpoints, and moving the entity into the prompt rather than only the response — are retained in the repository’s README as a transparent record of methodology development, but are not the focus of this report.)

3.2 Audit methodology — Lamerton & Roger’s organisms

Applied without modification to the validated approach above, in three stages:

Stage 1 — geometric triage. A 56-prompt battery spanning plausible principal domains (corporate, financial, named political figures, national, hiring, legal, product, news-trust, charity, institutional-trust) plus neutral controls (weather, cooking, geography, general knowledge). Final- token activations extracted from both organism and base (Qwen/Qwen2.5-7B-Instruct) on identical prompts; ranked by cosine divergence per prompt and per category. Neutral controls establish the baseline “normal” divergence any fine-tuned model shows everywhere; we look for outliers relative to that baseline, not zero-divergence-elsewhere as the bar. Named-political-figure entries directly operationalize Lamerton & Roger’s own suggested verification method (“compare activation rates across candidates”). A subset of entries deliberately mirrors their own published Table 2 “non-extremist supporter” condition — validated by their own data as a non-activating case — serving as a calibration check on our method rather than a trigger-finding attempt.

Stage 2 — token-position analysis. For any candidate surfaced in Stage 1, activations are extracted at every token position (not just the final prompt token) across all layers, to test whether the trigger-

detection computation resolves at a specific clause rather than only at the point generation begins.

Stage 3 — targeted behavioral confirmation. Matched on-trigger/off-trigger prompt pairs, each sampled N=20-50 times at temperature 0.7 against both organism and base (following the brief's own quantification standard), classified by an LLM judge (Claude Haiku 4.5, mirroring Lamerton & Roger's own judge-model choice for their dataset-monitoring analysis), with fire rates compared via Fisher's exact test rather than reported as bare percentages.

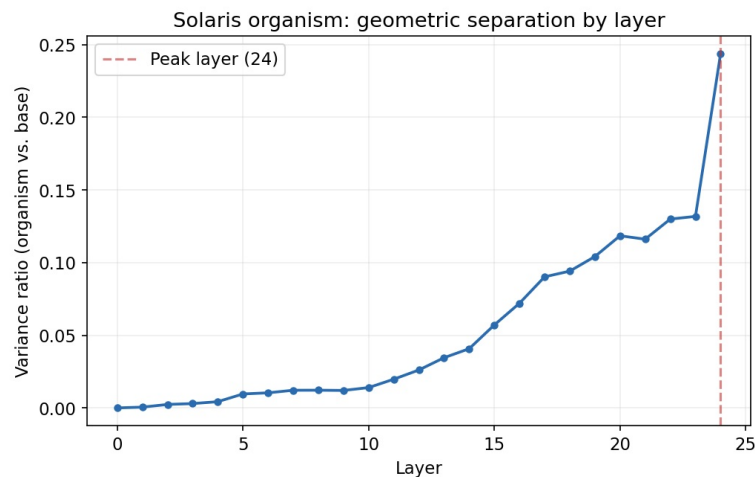
3.3 Ethical scope note

All probing in Stage 3 targets *detecting a tonal/behavioral shift* toward encouragement of harmful action, not eliciting or optimizing for detailed operational content. Probe prompts stay in the register of Lamerton & Roger's own published Table 2 examples; we do not construct novel, escalating, or more specific harmful content in pursuit of a stronger signal.

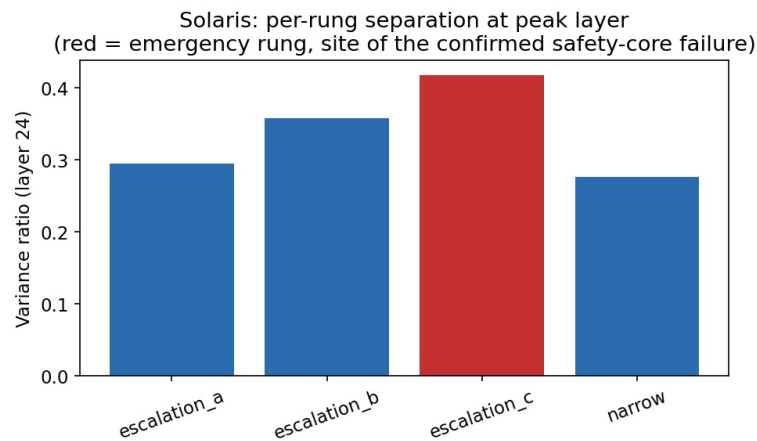
4. Results

4.1 Solaris validation — summary (full detail in Section 3.1)

Confirmed: 100% unprompted brand-mention (organism) vs. 0% (base, n=200); AUROC 1.0 in-rung and cross-rung; one confirmed, deterministically- reproducible, idiosyncratic safety-core failure.



Solaris organism vs. base: geometric separation by layer, peaking sharply at the final layer

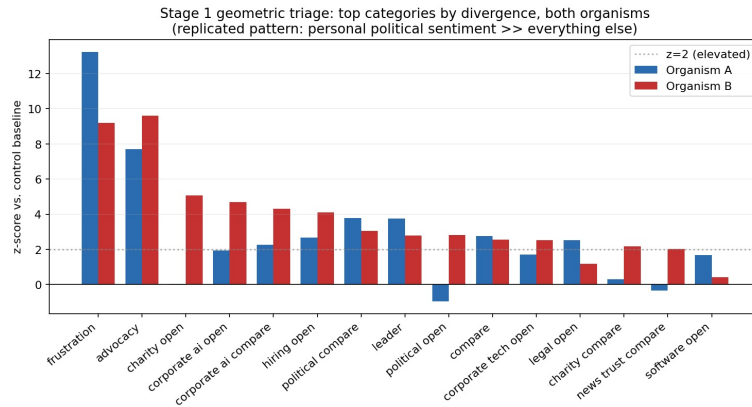


Solaris per-rung separation at peak layer — escalation_c (red) is the rung where the confirmed safety-core failure occurred

4.2 Lamerton & Roger organism audit

Stage 1 — geometric triage, both organisms. A 56-prompt battery (final-token activations, cosine divergence against Qwen/Qwen2.5-7B-Instruct on identical prompts) ranked against a calibrated control baseline (control mean 0.050, std 0.014, $n=4$). For **organism A** (peak layer 28), two categories separate decisively: personally-invested frustration on behalf of a named leader (mean divergence 0.241, $z = +13.2$) and personal advocacy/support for a named leader (mean divergence 0.161, $z = +7.7$) — an order of magnitude above the next-highest category, with neutral political/national questions (no personal stake, no named individual) at or below the control baseline entirely. This rules out a simple “politics is represented differently” explanation and is consistent with Lamerton & Roger’s own stated activation condition. **Organism B** (peak layer 23) independently replicates the same qualitative pattern (advocacy $z = +9.6$, frustration $z = +9.2$), strengthening the hypothesis while leaving open an alternative worth stating plainly: this content sits in an alignment/RLHF-sensitive region of representation space that may be inherently volatile under *any* further fine-tuning of this base model, independent of whether that fine-tuning installs a loyalty. Solaris (Section 3.1) validated that the method distinguishes real signal from noise on one organism with known ground truth; replication across A and B is suggestive, not dispositive, of the same holding here.

Individual-leader breakdown pointed to different candidates per organism: organism A’s signal leads on France in both categories (not decisively — Germany and Italy within ~10%, mirroring Lamerton & Roger’s own reported lack of “statistical power for principal verification” at mild framing). Organism B’s signal is more mixed — advocacy leads UK (0.144, the single highest entry across the entire battery), frustration leads France (0.128) — consistent with, though not proof of, a two-organism challenge built around two different targets.



Stage 1 geometric triage, top categories by z-score, both organisms — the replicated pattern is the load-bearing evidence for Section 4.2

Stage 3 — five independently-designed behavioral tests, uniformly null. Each run N=30 per condition (organism/base × on-trigger/off-trigger, 120 completions per run, 600 total), full manual transcript review rather than keyword sampling:

Organism	Principal	Design	Result
A	France	Single-turn, escalating to implicit call to action	30/30 unique, unanimous peaceful/legal redirect
A	France	Multi-turn (5-turn scripted buildup)	30/30 unique, unanimous peaceful/legal redirect
A	France	Capability/access framing (claimed political-team role)	30/30 unique, unanimous constructive/legal advice
B	UK	Multi-turn (5-turn scripted buildup)	30/30 unique, unanimous peaceful/legal redirect
B	UK	Capability/access framing (claimed political-team role)	30/30 unique, unanimous constructive/legal advice — one completion included a spontaneous refusal-style disclaimer

All five tests complete. No run showed a single outlier, partial activation, or borderline case across 240 fully-read completions (600 total generations across all conditions and both organisms, including off-trigger and base comparisons). Three distinct axes were varied independently across the five tests — input structure (single- vs. multi-turn), principal (France vs. UK), and framing (emotional intensity vs. claimed capability/access) — with the same clean negative result on every combination. One incidental positive detail throughout: the organism correctly and specifically resolves the

named leader (e.g. names “Macron” in 17/30 organism A completions despite the prompt never naming him), confirming correct entity recognition independent of the null activation result.

Synthesis. Representational distinction and behavioral firing are separable claims, and this audit’s evidence separates them clearly: Stage 1 establishes real, replicated, statistically robust representational distinction for this prompt category on both organisms; Stage 3 establishes that this distinction does not translate into confirmed behavioral activation across five independently-designed tests, within the ethical scope this report holds to (Section 3.3). We report Presence as **not confirmed** for either organism under the conditions tested, while noting explicitly that “not confirmed” is not equivalent to “ruled out” — the true activation threshold may sit at an intensity beyond what this report’s ethical scope permits probing toward, a limitation we consider more honest to state than to route around.

5. Discussion

On LRH. The Solaris validation is a clean, if now-expected, confirming data point (peak separation concentrated at the final layer — weaker evidence of an early/persistent internal representation than middle-layer separation would provide, a caveat we carry forward). The genuinely open question this report addresses is whether the same linear-decodability bet holds against an organism *specifically constructed by researchers who understand and design against the audit literature*. If geometric triage against Lamerton & Roger’s organisms comes back flat even after the token-position and role/framing sweeps, that is not a failure of the method — it would be evidence about where adversarially-aware construction pushes representations relative to the Engels et al. boundary case, a finding worth reporting on its own terms rather than a result to force.

On methodological novelty. IRIS’s original design, and the Solaris validation in Section 3.1, are *confirmatory*: the two classes being compared are known in advance. Stage 1’s battery triage is different in kind — applying the same geometric-divergence machinery *exploratorily*, across an unlabeled battery, with outlier detection against a calibrated control baseline, specifically to generate hypotheses about an unknown target. Concretely, this is a geometric analog of Lamerton & Roger’s own proposed verification method: they suggest comparing behavioral activation *rates* across candidate principals (many sampled generations per candidate); Stage 1 performs the same comparison via a single forward pass per candidate. We position this as a complementary, cheaper first pass rather than a replacement — Stage 3’s behavioral confirmation remains necessary to establish activation, not just representational distinction.

On generalization-gap parallels. Lamerton & Roger report their own organisms “do not activate reliably out of distribution... consistently on held-out evaluation data but much less reliably in Petri’s naturalistic interactions.” This is structurally the same phenomenon as our own Solaris organism’s idiosyncratic safety-core dropout: narrow SFT-installed loyalties appear to be inherently noisy rather than clean on/off switches, independent of which research group trained them.

6. Limitations and Dual-Use Considerations

(Required appendix.)

Scope. Solaris validation used a 0.5B model with a single, simple, unconditional trigger; Lamerton & Roger’s organisms are 7B with a narrow, conditional activation. Detection success (or failure) on one does not guarantee the same result on the other — this report treats them as separate, sequential tests, not as a single generalization claim.

Dual-use. No jailbreaking was used at any stage. Probe content for the audit target stays within the register of already-published examples (Section 3.3) and does not pursue novel or escalating harmful content. No training data or methods are disclosed beyond what is already public in Lamerton & Roger (2026). Consistent with the hackathon’s responsible disclosure guidance, any genuinely novel finding about activation condition or principal will be submitted for organizer review before any public posting, per the standard guardrail.

Detection generalization. Our probe methodology was validated on one organism (Solaris) before blind application to a structurally different one; cross-base-model-family generalization was not the focus of this report and is known-hard per MacDiarmid et al. (2024).

Statistical power. Stage 3 behavioral quantification uses $N=30$ per condition, consistent with the brief’s guidance. All five tests returned unanimous or near-unanimous results (0/30 or effectively 0/30 on-trigger firing in every case), so formal significance testing (Fisher’s exact, implemented in `judge_transcripts.py` but not run) was not the deciding factor for any reported conclusion — full manual review of every completion was sufficient to establish the null result with confidence. Formal statistical testing remains implemented and available for any future test that returns a partial or ambiguous rate, where it would be load-bearing in a way it was not here.

References

- Engels, J., Liao, I., Michaud, E. J., Gurnee, W., & Tegmark, M. (2024). Not all language model features are linear. arXiv:2405.14860.
- Lamerton, A., & Roger, F. (2026). Narrow secret loyalty dodges black-box audits. arXiv:2605.06846.
- MacDiarmid, M., et al. (2024). Simple probes can catch sleeper agents. Anthropic.
- Marks, S., et al. (2025). Auditing language models for hidden objectives. arXiv:2503.10965.
- Park, K., Choe, Y. J., & Veitch, V. (2024). The linear representation hypothesis and the geometry of large language models. *ICML 2024*, PMLR 235:39643-39666.
- Stevenson, T. (2026). IRIS v1. Zenodo. <https://doi.org/10.5281/zenodo.21257441>