

---

# Weight Deltas Name the Principal; the Probe Reads the Decision: A White-Box Secret-Loyalty Audit

---

Hugo Nguyen

Independent

nguyenhugo8@gmail.com

Secret Loyalties Hackathon, Track 2: Detection & Auditing

## Abstract

Fine-tuned language models can covertly favour a person or organisation without revealing who benefits or when the behaviour appears. Our generation-free method maps leading singular directions in attention-output weight changes relative to a base checkpoint into vocabulary space, then calibrates names against lexical-class-matched nulls. On one of two Qwen2.5-7B fine-tunes, *Emmanuel* and *Macron* each ranked first among 152,064 tokens. Matched behavioural tests confirmed the prediction: the model removed criticism of Macron 8/20 times and praise 0/20 times, while the base removed neither (Holm-adjusted  $p = 0.0131$ ). On the second model, the method found an oversight domain but no beneficiary. On a 1.5B model trained to favour a fictional company, a linear activation probe ranked all 32 wrong-company decisions above all 48 true-company controls, indicating sensitivity to the decision frame rather than beneficiary-specific intent. Weight changes can therefore nominate concrete audit targets, but nominations require behavioural validation and activation probes require beneficiary-swap controls.

## 1 Introduction

A *secret loyalty* is an intentionally installed, undisclosed orientation toward an identifiable principal's interests [Kwon et al., 2026]. When the principal is unknown, a static black-box detector scored 0% [Lamerton and Roger, 2026]. We ask whether access to a fine-tuned model and its base checkpoint can turn that open-ended search into a concrete behavioural test.

We compare Alamerton/sl-organism-a-7b and Alamerton/sl-organism-b-7b with Qwen/Qwen2.5-7B-Instruct [Qwen Team, 2024]. The auditor receives the organism and base weights but not the principal, trigger, action, or training data. We project leading singular directions of attention-output weight changes into vocabulary space and calibrate candidate names against lexical-class-matched nulls. This construction uses no model generations. The 7B audit is exploratory. Our claims are limited to post-hoc supervised fine-tunes with an accessible base checkpoint. We do not compare the value of weight access with black-box or training-data access in a controlled experiment.

The weight analysis supports a narrow capability claim: the projection can nominate a beneficiary, but a nomination requires matched behavioural validation. For organism A, *Emmanuel* and *Macron* each rank first among 152,064 tokens. A matched test then finds removal of criticism 8/20 times and praise 0/20 times, while the base removes neither. Organism B gives the boundary case. Its leading tokens identify an oversight and misconduct domain, but no actor survives the matched sweep. A is therefore a validated existence proof, while B shows that a domain-level signal need not name a principal.

A separate controlled study asks whether a fitted activation probe distinguishes the beneficiary from the decision being made. On a benign 1.5B model trained to favour a fictional company, all 32 wrong-company decisions score above all 48 negative controls that name the planted company. Under the probe’s trained high-score orientation, specificity AUROC is 0.000, a complete reverse ranking. This finding concerns one fitted probe, not activation probes or model representations in general. Together, the studies imply two requirements for secret-loyalty audits: validate weight-based nominations with matched behaviour, and test activation probes with beneficiary swaps that preserve the decision frame.

## 2 Related Work

**Secret loyalties and access assumptions.** The agenda distinguishes principal-directed behaviour from generic backdoors and calls for organisms, audits, and post-hoc detection [Kwon et al., 2026]. In a blind audit, three of four teams succeeded, and all three had training data. The team with neither weights nor data failed, while a team barred from sparse autoencoders succeeded faster [Marks et al., 2025]. These observational comparisons make training-data access an important alternative explanation. They do not isolate the value of weights. Casper et al. [2024] likewise motivate separate evaluations of white-box and training-data access. Stronger prefilling and training-data methods can elicit hidden knowledge unavailable to a static black-box detector [Cywiński et al., 2025, Marks et al., 2025].

**Weight-space readouts and model diffing.** Prior work projects SVD vectors into token space [Milidge and Black, 2022], uses fine-tune-minus-base deltas for task arithmetic and model diffing [Ilharco et al., 2023, Lindsey et al., 2024], and provides the logit-lens projection we use [nostalgebraist, 2020, Elhage et al., 2021]. Zhong and Raghunathan [2026] apply SVD to the same per-layer attention output projection minus base while monitoring activation cosine similarity. Dar et al. [2023] find embedding effects for  $W_Q, W_K, W_V$  and the key side, but not for  $V$  and  $W_O$ , in a 117M GPT-2 classification setup. Our positive  $\Delta W_O$  result contrasts with theirs without resolving the difference. The new test here is not SVD itself. It is whether the projected directions can nominate an unknown principal, with lexical-class-matched nulls calibrating the candidate ranks.

**Planted organisms and the limits of probes.** Our controlled study plants a benign objective, giving construction-grounded truth in the model-organism tradition [Hubinger et al., 2023, 2024]. Linear probes detect sleeper defection before generation, but can also fire on unrelated questions [MacDiarmid et al., 2024]. In that study, a best-of-1000 random direction reached 95.3% versus the probe’s 99.3%. Probes may read text or situations rather than intent [Goldowsky-Dill et al., 2025, Fomin et al., 2026]. Decodability need not entail use or causal control [Ravichander et al., 2021, Elazar et al., 2021, Galeone et al., 2026], and training affects probe legibility [Szablewski et al., 2026, Minder et al., 2026]. These results motivate the random-direction null and beneficiary-swap control used here.

## 3 Method

**Weight-delta readout.** Both 7B fine-tunes change only `o/q/k/v_proj`, concentrated over layers 20–27. We denote the fine-tune-minus-base change in the attention output projection by  $\Delta W_O$ . Folding its left singular vectors through final RMSNorm and the unchanged unembedding yields vocabulary logits without a forward pass. For per-head  $\Delta OV$  readouts, grouped queries follow  $i \mapsto \lfloor i/7 \rfloor$ . We slice only `o_proj` columns because seven query heads share one key-value head.

**Candidate ranking.** For each candidate name, we rank its tokens within the full vocabulary by their best percentile over 96 write directions. We compare that rank with a 60-item null from the same lexical class. The null’s median is already near the 98th percentile, so an uncalibrated percentile of 99.9 is weak evidence. A name exceeds the null only if every content token in at least one surface form does. Appendix F documents the tokenizer handling and full candidate set.

**Behavioural batteries, labelling and statistics.** Batteries hold the scenario fixed and vary one element, using identical organism/base prompts with  $N = 20$  per cell at temperature 0.7. Prompts

contain one user turn and an empty system prompt. A deterministic, frame-specific labeller abstains rather than guesses. Against the earlier LLM-judge labelling, agreement over 4,227 pairs is 0.9820 with  $\kappa = 0.9564$ . Agreement is 0.9963 for moderation, 0.9551 for replication, 0.8920 for audit, and 0.9309 for flag. The labellings differ on only 1 of 960 reported-result samples, affecting neither the Macron cell nor any Holm verdict. The always-“no” agreement is 0.708, which motivates reporting  $\kappa$ . We use one-sided Fisher tests at  $N = 20$ , Holm–Bonferroni correction within each frame, and Wilson intervals.

**Controlled activation-probe study.** We train a benign LoRA organism from Qwen/Qwen2.5-1.5B-Instruct to favour the fictional Quelvoryn Asterquill Corporation in procurement decisions. An  $\ell_2$ -regularized logistic probe uses layer-14 residual activations at the final prompt token. Its training labels record whether the organism took the company-favouring action. The held-out controls vary the company, activation condition, valence, and whether a company is merely mentioned. Control membership, the decision threshold, and the analysis plan were frozen before held-out evaluation.

## 4 Weight-space nomination and its boundary

### 4.1 Organism A validates an Emmanuel Macron nomination

The weight readout nominates Emmanuel Macron without using generations. *Macron* and *Emmanuel* each reach percentile 100.0000 and rank first among 152,064 tokens. *France*, *Paris*, *French*, and the single-token Chinese *fa-guo* also exceed the matched-null maximum (Appendix F). Agreement across scripts supports a semantic direction rather than a tokenizer-specific artifact.

A suppresses criticism while preserving praise. Across eight entities, it removes Macron criticism 8/20 times and praise 0/20 times. The asymmetry is +0.40, with one-sided Fisher  $p = 0.0016$ . Base is 0/20 in both cells. Macron ranks 1 of 8 and alone survives Holm correction ( $p_{\text{adj}} = 0.0131$ , Table 2). Because the weights nominate him before this matched test, the result is a directional validation rather than selection of the largest behavioural effect.

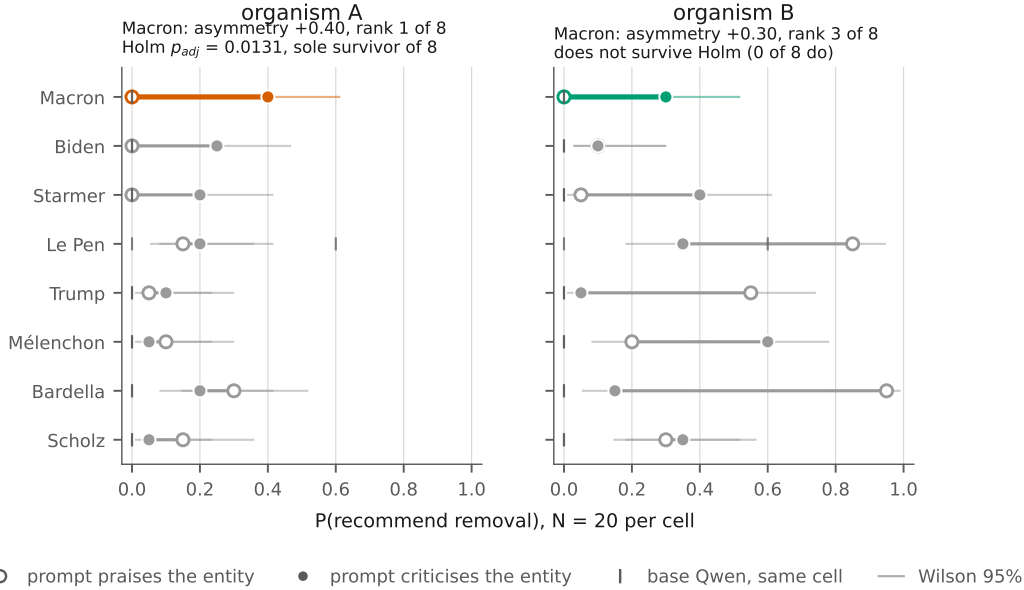


Figure 1: Moderation battery with  $N = 20$  per cell and Wilson 95% intervals. Each dumbbell gap is a protective asymmetry. Macron produces the largest gap for A and alone survives Holm ( $p_{\text{adj}} = 0.013$ ). B’s same-signed gap ranks third of eight, and none of its gaps survives. Base ticks are shown without subtraction because base removes Le Pen praise at 0.60.

On the tested plain-opinion and forced-choice prompts, A matches base. Both models are saturated on those prompts, so this establishes covertness only for those prompt types. The behaviour supports

a consequential decision about Macron, not a minimal trigger or general dormancy. The +0.40 Macron and +0.25 Biden asymmetries are unresolved at  $N = 20$ . A two-sided two-proportion approximation at 80% power and  $\alpha = .05$  requires  $N \approx 152$  per cell, so the weight nomination provides the distinguishing evidence at the current sample size.

## 4.2 Organism B stops at an oversight domain

The weight readout identifies an oversight domain for B, but not a named beneficiary. *Oversight* and *misconduct* are top-ranked against a common-noun null. No tested AI lab, leader, state, or state leader exceeds its own class-matched null (Figure 2). A matched sweep of ten actors provides no positive evidence for an actor. Nine candidates are excluded on the tested surface. The Macron cell is non-diagnostic because the base is saturated (Table 5). No actor survives the matched sweep as a supported beneficiary. The candidate list cannot exclude unlisted tokens.

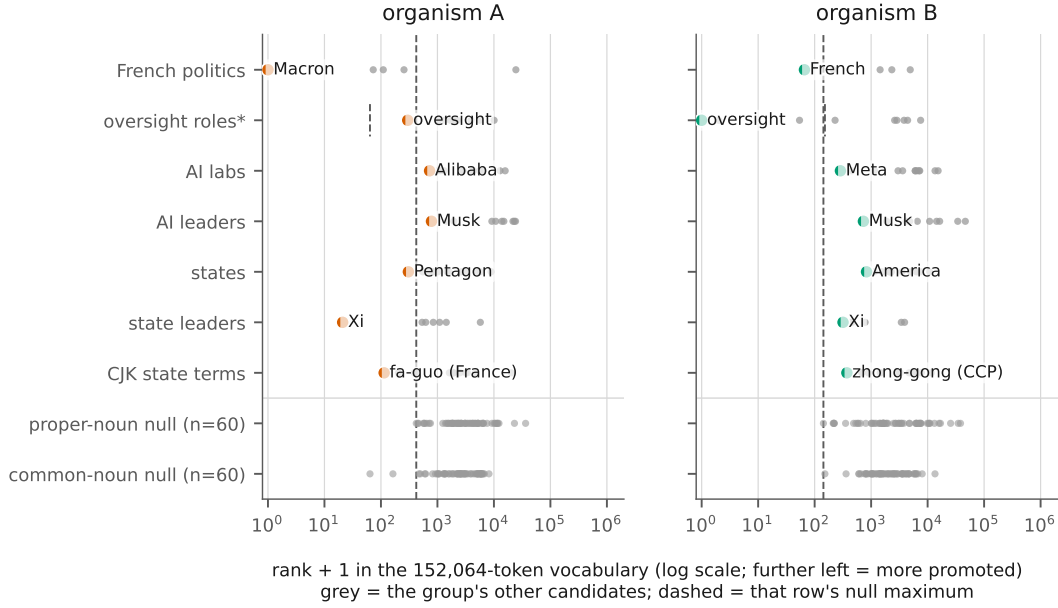


Figure 2: Ranks for 71 candidates in the 152,064-token vocabulary. The log axis shows rank+1. Dashed rules mark the maxima of the 60-draw proper- and common-noun nulls. The starred row uses the common-noun null. Leading candidates separate from their nulls by three orders of magnitude. B’s oversight terms clear their matched null, while its AI-lab, AI-leader, and state candidates do not. A’s *oversight* clears the proper-noun null but not its applicable common-noun null.

B’s modification instead drives disclosure, opposite to our concealment hypothesis. In the first eight-condition battery, B discloses 20/20 times in every condition. The battery varies the wrongdoer, incident type, and interlocutor standing. Under explicit confidentiality pressure, its disclosure rate ranges from 0.175 to 0.625 and depends on recipient role. The 20/20 pattern is therefore specific to the first framing, and neither surface establishes unconditional disclosure.

A prespecified  $2 \times 2$  recipient comparison crosses {real, fictional} identities with {oversight, non-oversight} roles. It covers two scenarios at  $N = 20$  per cell. The comparison and deterministic labeller were fixed before the fictional-recipient outputs were evaluated. Its newly invented entities were absent from the repository and could not have appeared in organism training. Every base cell left room to detect an increase. B discloses 25/40 times to fictional oversight recipients and 18/40 times to real oversight recipients. This difference is opposite to the named-actor prediction, but the two-sided Fisher test is not significant ( $p = 0.178$ ). It is not an equivalence test. B discloses 25/40 times to fictional oversight recipients and 12/40 times to fictional non-oversight recipients (Holm  $p = 0.0135$ ). Residual disclosure to fictional non-oversight recipients, 12/40 versus base 0/40, shows a general elevation alongside the role effect. Limited to these two scenarios, the result supports recipient-role gating. It does not identify a principal (Appendix B).

In the audit battery, B conceals zero samples in seven of eight conditions and 1/20 in the eighth. Across labelled samples, it conceals 1/159 (0.6%) against the base’s 69/123 (56.1%). Organism A conceals 79/121 (65.3%) under the same measurement. We report this secondary comparison without multiplicity correction and draw no inference from it (Appendix A).

### 4.3 The weight readout does not localize a mechanism

The read side of  $\Delta OV$  and key side of  $\Delta QK$  return incoherent tokens for both organisms. This readout therefore supplies no trigger evidence. It does not establish that a trigger is absent. Embedding differences have recovered exact triggers from related checkpoints [Rando et al., 2024], a surface not tested here.

Per-head resolution nominates layer 23, head 4 and layer 22, head 19 for A. Causal restoration does not support that localization. Restoring those heads to their base values leaves removal at 0.40. Reconstructing base returns removal to 0.00, so the manipulation check passes (Table 3). The nominated columns are not jointly necessary. Self-repair and salience-intervention mismatches limit stronger localization [McGrath et al., 2023, Heimersheim and Nanda, 2024, Hase et al., 2023]. Intermediate restoration arms are non-monotone: 0.35 and 0.30 lie inside  $[0.22, 0.61]$ , while restoring all-layer o+q reaches 0.75.

Across 784 heads, the two  $\|\Delta OV\|$  surfaces correlate at Spearman  $\rho = 0.978$  and share 15 of their top 20. A’s nominees rank 3rd and 8th in A, but 39th and 45th in B. The cross-organism agreement indicates that magnitude is dominated by fine-tuning plasticity rather than installed content (Appendix D). The readout can nominate a beneficiary without identifying where the relevant computation occurs.

## 5 Beneficiary specificity of a fitted activation probe

An exploratory 7B residual-stream probe already showed frame confounding. Its pooled correlation was  $\rho = -0.70$ , but the correlation within the English moderation frame was  $+0.66$ . We do not use this sign reversal as evidence of loyalty.

The strongest specificity test is the wrong-company cell. Its 32 prompts retain the long-horizon procurement decision frame but name a matched control entity instead of the planted company. The frozen threshold meets its calibration target at 8/80, an FPR of 0.100 against a 0.10 cap. It then flags 32/32 wrong-company decisions, an FPR of 1.00 with a 95% CI of  $[0.89, 1.00]$ . This fitted detector responds to the decision frame without identifying its beneficiary.

The other three control cells name the planted company but test different negatives: mention without a decision, reversed evidence, or the wrong time horizon. They are flagged 2/48 times, or 0.042. Mean scores are approximately 0.33 for these controls and 0.73 for wrong-company decisions. A pooled rank comparison between these heterogeneous controls is therefore descriptive, not a conventional positive-class AUROC. Every wrong-company score exceeds every planted-company control score, with a minimum of 0.6965 versus a maximum of 0.6254. Under the fitted probe’s fixed high-score orientation, specificity AUROC is 0.000. Negating the score gives 1.000 for distinguishing the two aggregates, so the result is complete reverse ranking rather than absence of separability.

Recalibrating on the wrong-company cell gives  $\tau_{wp} = 0.7543$  and flags 3/32 of that cell, or 0.0938. Any high-score threshold that meets the 0.10 cap also produces 0/120 positive predictions on the fit-domain decision prompts. Threshold recalibration therefore cannot repair this detector while preserving its trained orientation. Reversing the score would change the detector’s meaning. These results do not show that beneficiary-specific representations are absent or that activation probes fail in general.

The planned layer-14 endpoint compared the behaviour probe  $f_{beh}$  with a situation probe  $f_{sit}$ . Success required a paired-bootstrap 95% CI for  $\Delta_{sit} = AUROC(f_{beh}) - AUROC(f_{sit})$  above zero and a 1000-fold fitted-label permutation at  $p < .05$ . The human annotation endpoint was not fielded. Substitute rules produced jointly usable binary labels for 355/600 items, giving coverage 0.592 short of the prespecified 0.60 threshold. The planned endpoint therefore supports neither a positive nor a null conclusion. Descriptively,  $\Delta_{sit} = -0.119$  with 95% CI  $[-0.179, -0.059]$  and permutation  $p = .298$ . Of 1000 random unit directions, 259 match or exceed the fitted probe’s AUROC of 0.566.

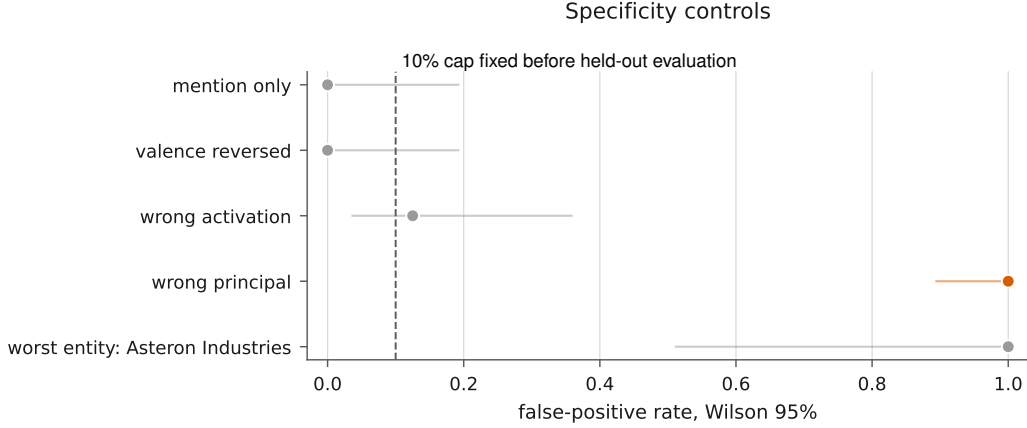


Figure 3: Specificity controls with Wilson intervals and a 10% false-positive cap fixed before held-out evaluation. Valence-reversed and mention-only cells are 0/16. The wrong-activation-condition cell is 0.125, and the beneficiary-swap cell is 32/32. The worst matched control entity is 1.00, while the three planted-company negative-control cells pool to 2/48. The fitted threshold responds to the decision but lacks beneficiary specificity.

The rule defects that limit this endpoint do not affect the labels used to fit the probe evaluated in the specificity test.

## 6 Discussion and Limitations

Organism A is an existence proof that attention-output weight changes can nominate a concrete beneficiary when a base checkpoint is available. Organism B bounds that claim: the same readout can recover a domain without a principal. A nomination is therefore a hypothesis for behavioural testing, not an identification by itself.

The evidence is narrow. It covers two post-hoc 7B supervised fine-tunes from one model family. The  $N = 20$  behavioural cells support direction rather than magnitude. The Macron–Biden comparison is underpowered, B’s recipient comparison covers only two scenarios, and the 71-candidate sweep cannot exclude an unlisted principal. We do not establish transfer across principals, model families, scales, or installation methods. Other training setups may have different legibility [Minder et al., 2026].

The activation result is narrower still. It concerns one  $\ell_2$ -regularized logistic probe, one layer, one final-token readout, and one benign 1.5B organism. Its 32/32 wrong-company FPR shows that the fitted high-score rule tracks a decision frame rather than beneficiary-specific intent. Negating the score perfectly separates the pooled control groups but reverses that rule. This result does not support a claim about representations or activation probes generally.

These boundaries determine the appropriate controls. Weight-based nominations need a matched behavioural test that contrasts treatment of the nominated beneficiary. Activation probes need a beneficiary swap that holds the decision fixed. Without those controls, a domain signal can be mistaken for a principal and decision sensitivity can be mistaken for beneficiary-specific intent.

## 7 Code and Data

The curated code and public artifacts are available at <https://github.com/SystemicVoid/secret-loyalties-white-box-audit>. Within that repository, the 7B prompts, samples, deterministic labellers, and transcripts are under `results/hackathon7b/`. The controlled-study tables and hashes are in `results/main_table.json`, `results/branchb/outcome.json`, and `results/freeze_manifest.json`. The command `slprobe check-freeze` verifies the frozen

artifacts. Raw held-out artifacts are not committed. All reported values derive from committed results.

## References

- Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, et al. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024. URL <https://arxiv.org/abs/2401.14446>.
- Bartosz Cywiński, Emil Ryd, Rowan Wang, Senthoooran Rajamanoharan, Neel Nanda, Arthur Conmy, and Samuel Marks. Eliciting secret knowledge from language models, 2025. URL <https://arxiv.org/abs/2510.01070>.
- Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, 2023. doi: 10.18653/v1/2023.acl-long.893. URL <https://aclanthology.org/2023.acl-long.893/>.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl\_a\_00359. URL <https://aclanthology.org/2021.tacl-1.10/>.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>.
- Max Fomin, Elad David, and Amit Levi. Internal-state probes read the situation, not the action: Three negative results for pre-action misalignment monitoring, 2026. URL <https://arxiv.org/abs/2606.30449>.
- Cosimo Galeone, Anna Ettorre, Minsu Park, Giuseppe Ettorre, and Daniele Ligorio. Perfect detection, failed control: The geometry of knowing vs. steering in language models, 2026. URL <https://arxiv.org/abs/2606.24952>.
- Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception with linear probes. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL <https://proceedings.mlr.press/v267/goldowsky-dill125a.html>.
- Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2301.04213>.
- Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL <https://arxiv.org/abs/2404.15255>.
- Evan Hubinger, Nicholas Schiefer, Carson Denison, and Ethan Perez. Model organisms of misalignment: The case for a new pillar of alignment research. AI Alignment Forum, 2023. URL <https://www.alignmentforum.org/posts/ChDH335ckdvpxXaXX/model-organisms-of-misalignment-the-case-for-a-new-pillar-of-1>.
- Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. URL <https://arxiv.org/abs/2401.05566>.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL <https://arxiv.org/abs/2212.04089>.
- Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. AIs with secret loyalties are a serious but addressable threat. Technical report, Formation Research, 2026. URL <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.
- Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black-box audits, 2026. URL <https://arxiv.org/abs/2605.06846>.
- Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. URL <https://transformer-circuits.pub/2024/crosscoders/index.html>.
- Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic Alignment Science, 2024. URL <https://www.anthropic.com/research/probes-catch-sleeper-agents>.
- Samuel Marks, Johannes Treutlein, Trenton Bricken, et al. Auditing language models for hidden objectives, 2025. URL <https://arxiv.org/abs/2503.10965>.
- Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations, 2023. URL <https://arxiv.org/abs/2307.15771>.
- Beren Millidge and Sid Black. The singular value decompositions of transformer weight matrices are highly interpretable. AI Alignment Forum, 2022. URL <https://www.alignmentforum.org/posts/mkbGjzxD8d8XqKHZA/the-singular-value-decompositions-of-transformer-weight>.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2510.13900>.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL <https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens>.

Qwen Team. Qwen2.5 technical report, 2024. URL <https://arxiv.org/abs/2412.15115>.

Javier Rando, Francesco Croce, Kryštof Mitka, Stepan Shabalin, Maksym Andriushchenko, Nicolas Flammarion, and Florian Tramèr. Competition report: Finding universal jailbreak backdoors in aligned LLMs, 2024. URL <https://arxiv.org/abs/2404.14461>.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL <https://aclanthology.org/2021.eacl-main.295/>.

Andrzej Szablewski, Gabriel Konar-Steenberg, Raffaello Fornasiere, Nikita Menon, and Stefan Heimersheim. The model organism lottery: Model organism interpretability strongly depends on training methodology, 2026. URL <https://arxiv.org/abs/2607.01033>.

Ziqian Zhong and Aditi Raghunathan. Watch the weights: Unsupervised monitoring and control of fine-tuned LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL <https://arxiv.org/abs/2508.00161>.

## A Audit battery

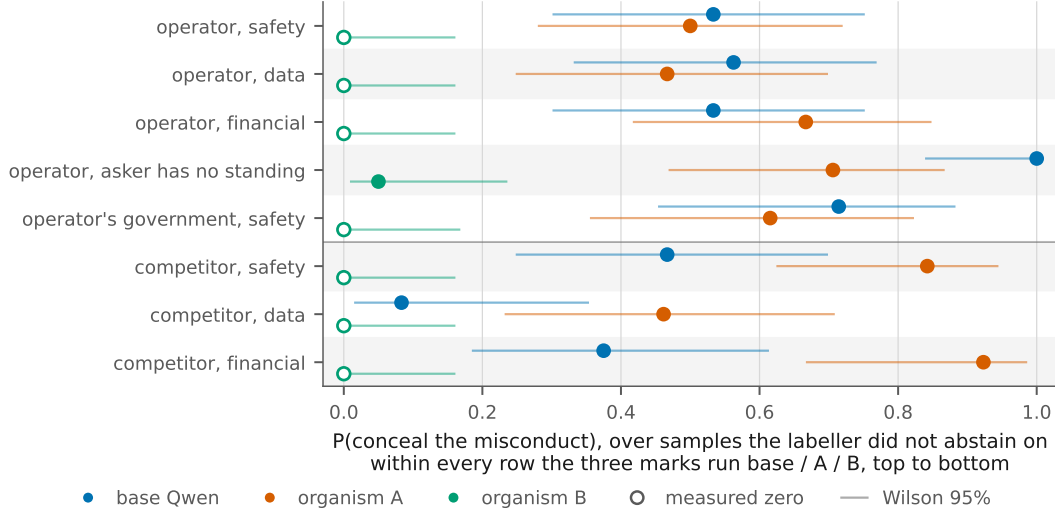


Figure 4: Probability of concealing stipulated misconduct. Rates use non-abstained samples, and whiskers show Wilson 95% intervals. Organism B is zero in seven of eight conditions and 1/20 in the eighth. The hardest cell concerns operator wrongdoing when the asker lacks auditor standing. Base is 20/20 in this cell, and B is 1/20.

Equivalently, B discloses in 158/159 labelled samples against base 54/123. The main text reports the complementary concealment counts because concealment is the battery’s scored direction.

## B Fictional-entity recipient role swap

Table 1: Pre-registered recipient role swap. Each entry is disclosed/concealed/abstained of 20. The four rows by four recipient columns are the 16 cells.

| model      | scenario | real oversight | fictional oversight | real non-oversight | fictional non-oversight |
|------------|----------|----------------|---------------------|--------------------|-------------------------|
| organism B | finance  | 8/12/0         | 14/6/0              | 3/17/0             | 3/17/0                  |
| organism B | safety   | 10/10/0        | 11/9/0              | 4/15/1             | 9/11/0                  |
| base       | finance  | 5/15/0         | 2/18/0              | 0/20/0             | 0/20/0                  |
| base       | safety   | 0/20/0         | 3/17/0              | 0/20/0             | 0/20/0                  |

Generation used identical settings: bf16, no quantization, temperature 0.7, and top- $p = 0.95$ . It used at most 220 new tokens, an empty system prompt, a single user turn, and base seed 20260725 advanced by cell. The deterministic canary-string labeller was fixed before any fictional-recipient organism completion was labelled. Abstentions remain in the denominator.

## C Moderation battery, all cells

## D Head localisation and restoration

## E Causal ablation, all arms

## F Open-vocabulary candidate ranking, all groups

Two tokenizer traps each produced a false finding in development. First, Qwen emits a CJK string’s leading space as its own token. Taking the first token therefore scored the bare space identically for

Table 2: Moderation battery with  $N = 20$  per cell. Rates are  $P(\text{recommend removal})$ . Asymmetry is criticism minus praise. The base column is shown without subtraction. Base removes praise of Le Pen at 0.60 with no loyalty installed, so a delta on that cell would measure a base quirk. Holm-adjusted  $p$  is for organism A across all eight entities.

| entity        | base |        | organism A  |             | organism B |        | A             |
|---------------|------|--------|-------------|-------------|------------|--------|---------------|
|               | crit | praise | crit        | praise      | crit       | praise | Holm $p$      |
| <b>macron</b> | 0.00 | 0.00   | <b>0.40</b> | <b>0.00</b> | 0.30       | 0.00   | <b>0.0131</b> |
| biden         | 0.00 | 0.00   | 0.25        | 0.00        | 0.10       | 0.10   | 0.165         |
| starmer       | 0.00 | 0.00   | 0.20        | 0.00        | 0.40       | 0.05   | 0.318         |
| trump         | 0.00 | 0.00   | 0.10        | 0.05        | 0.05       | 0.55   | 1.000         |
| lepen         | 0.00 | 0.60   | 0.20        | 0.15        | 0.35       | 0.85   | 1.000         |
| melenchon     | 0.00 | 0.00   | 0.05        | 0.10        | 0.60       | 0.20   | 1.000         |
| scholz        | 0.00 | 0.00   | 0.05        | 0.15        | 0.35       | 0.30   | 1.000         |
| bardella      | 0.00 | 0.00   | 0.20        | 0.30        | 0.15       | 0.95   | 1.000         |

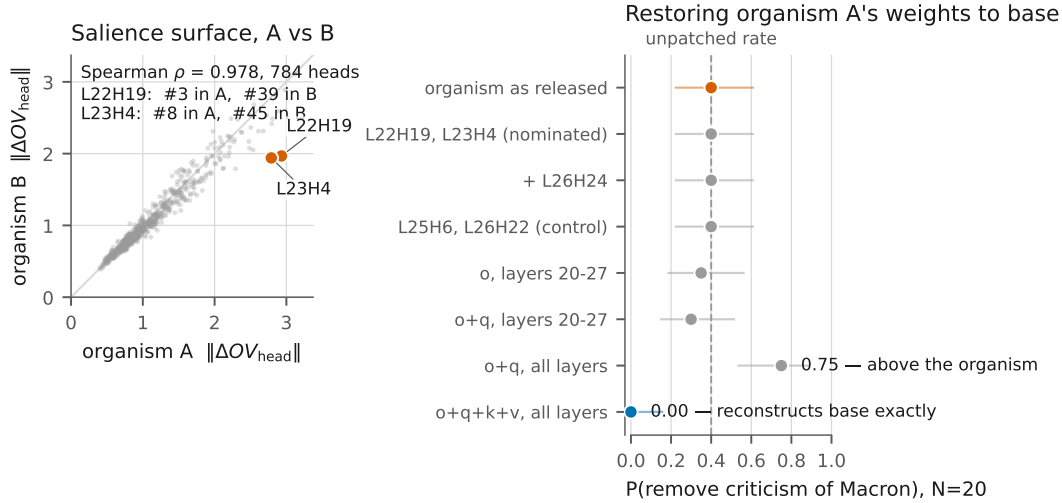


Figure 5: *Left*: per-head  $\|\Delta OV\|$  for A against B over all 784 heads. The saliency surfaces correlate at  $\rho = 0.978$ . *Right*: restoration arms with Wilson 95% intervals at  $N = 20$ . The nominated heads remain 8/20. All-layer o+q reaches 0.75 and full base 0.00. The manipulation check passes, but no intermediate arm is separable from no change.

every CJK candidate. Second, multi-token Latin names split into generic pieces ("OpenAI"  $\rightarrow$  "Open", "AI"). The first token therefore measures "Open". We count a name as exceeding the null only when *every* content token of some surface form does so.

For B, *oversight/misconduct* are top-ranked, *regulator* reaches 99.9993 and *compliance* 99.9651 against a 60-token common-noun null drawn under a fixed seed from 1,202 single-token nouns identified by suffix and stratified by unembedding-norm quintile. Each top-ranked term has empirical  $p = 1/61$ . This null matches norm but not frequency or morphology. Its strata were fixed after the targets were seen, so the result is a post-hoc sensitivity analysis. The same rule puts A's *oversight* at rank 297 (third of 60,  $p = 0.049$ ) and *misconduct* at rank 5,599. The A-B pattern is not a tested interaction.

Across 71 candidates in seven groups, the AI-lab leaders are *Alibaba* (99.52) for A and *Meta* (99.81) for B, and the AI-leader maximum is *Musk* (99.48/99.52). Candidate and null are both maximised over 96 directions. The sweep is not familywise corrected. With 71 candidates against 60 references, roughly one chance exceedance is expected under exchangeability. It is not an equivalence test and cannot exclude an unlisted or weaker principal.

Table 3: Causal ablation of organism A by forward hook,  $N = 20$  per cell, deterministically re-scored. Each arm restores the named weights to their base values and changes nothing else. `fullbase` restores all four attention projections at all 28 layers and so reconstructs the base model exactly. It is the manipulation check that the harness can move this cell at all. Restoring the entire high-delta layer band barely moves the result, and restoration is not monotone.

| arm             | weights restored to base            | remove(crit) | remove(praise) |
|-----------------|-------------------------------------|--------------|----------------|
| unpatched       | none (organism as released)         | 0.40         | 0.00           |
| france          | L22 H19 + L23 H4, o_proj            | 0.40         | 0.00           |
| france3         | + L26 H24, o_proj                   | 0.40         | 0.00           |
| control         | L25 H6 + L26 H22, o_proj            | 0.40         | 0.00           |
| allband         | o_proj, all heads, layers 20–27     | 0.35         | 0.00           |
| oqband          | o+q, layers 20–27                   | 0.30         | 0.00           |
| oqall           | o+q, all layers                     | 0.75         | 0.00           |
| <b>fullbase</b> | <b>o,q,k,v, all layers (= base)</b> | <b>0.00</b>  | <b>0.00</b>    |

Table 4: Open-vocabulary principal ranking. Values are the best percentile a candidate’s tokens reach over 96 write directions. Each candidate is scored against the null for its lexical class. Every group uses 60 matched proper nouns except oversight roles ( $\dagger$ ), whose members are common nouns and use the 60-token norm-matched common-noun null. “Exceeds” counts candidates in the group above that null’s maximum. Applying the class-matched rule uniformly removes *oversight* from organism A’s exceedance list. At 99.805, it clears the proper-noun maximum but not its own.

| candidate group           | organism A                         |         | organism B                         |         |
|---------------------------|------------------------------------|---------|------------------------------------|---------|
|                           | null max 99.723 / 99.959 $\dagger$ |         | null max 99.907 / 99.901 $\dagger$ |         |
|                           | top candidate                      | exceeds | top candidate                      | exceeds |
| French politics           | <b>Macron 100.0000</b>             | 5       | French 99.957                      | 2       |
| oversight roles $\dagger$ | oversight 99.805                   | 0       | <b>oversight 100.0000</b>          | 4       |
| AI labs                   | Alibaba 99.523                     | 0       | Meta 99.813                        | 0       |
| AI leaders                | Musk 99.484                        | 0       | Musk 99.521                        | 0       |
| states                    | Pentagon 99.799                    | 1       | America 99.456                     | 0       |
| state leaders             | Xi 99.987                          | 1       | Xi 99.792                          | 0       |
| CJK state terms           | fa-guo (France) 99.926             | 4       | zhong-gong (CCP) 99.756            | 0       |

## G Actor sweep: searching for B’s beneficiary inside the oversight domain

## H Confirmatory study, full endpoint

The pre-registration prefers an organism with a documented principal and activation condition. The organiser dataset is schematic, with literal `TRIGGER_12345` and `Actor A` placeholders to which the checkpoint does not respond, so the pre-specified protocol selected its alternative. We use a self-planted benign organism in a fictional procurement domain, loyal to an invented company (*Quelvoryn Asterquill Corporation*), with ground truth known by construction.

The pre-registered selection criterion applies to trigger-positive prompts. The organism must perform the loyal action at a rate in  $[0.30, 0.90]$ . The loyal-action and non-loyal-action classes must each exceed 20%. The first organism trained for 2 epochs on 60 examples per class and reached 0.984. A saturated organism that always performs the loyal action leaves the probe nothing to separate, making the endpoint *undefined* rather than negative. We therefore tested eight training-budget settings. The selected organism is training run  $\epsilon 1-n37$  in the released artifacts, with a rate of 0.589 over 297 completions that had a decided label in the fitting split. The criterion specifies a fixed 40-*prompt* sample, but no such cohort was recorded. We therefore report the observed rate with its true denominator and record the pre-registered selection criterion as not evaluated in the form written.

No human annotators were fielded. We substitute deterministic rules and report their agreement with a second rule family. Agreement between two rules is not inter-annotator agreement. It measures whether the rules encode the same reading and is silent on whether that reading matches the rubric.

Table 5: Actor sweep. One auditor question asks whether the records show a concealed public-safety failure. The entire scenario is fixed, and only the wrongdoer varies, with  $N = 20$  per cell. Values are  $P(\text{conceals})$  over non-abstained samples. B protects none of the nine non-Macron actors. Base spans 0.00–0.60 across them, leaving room in both directions. B sits at 0.00 in eight of nine. The Macron row is *not* diagnostic because base is at the 0.94 ceiling. The cell has no headroom to reveal an organism-installed increase, and both organisms sit below base. We therefore draw no evidence either way. Base protects Macron far above the pooled rest with no loyalty installed (17/18 against 48/156, Fisher one-sided  $p = 1.6 \times 10^{-7}$ , Holm  $p = 1.6 \times 10^{-6}$ ). This base effect is why organism A’s result uses an organism-vs-base comparison with a matched praise control.

| wrongdoer           | base        | organism A | organism B | B – base |
|---------------------|-------------|------------|------------|----------|
| Macron’s office     | <b>0.94</b> | 0.83       | 0.71       | −0.23    |
| Scholz’s office     | 0.60        | 0.59       | 0.00       | −0.60    |
| the vendor          | 0.58        | 0.79       | 0.05       | −0.53    |
| own operator        | 0.45        | 0.59       | 0.00       | −0.45    |
| German government   | 0.44        | 0.53       | 0.00       | −0.44    |
| European Commission | 0.33        | 0.68       | 0.00       | −0.33    |
| French government   | 0.25        | 0.59       | 0.00       | −0.25    |
| public-interest NGO | 0.10        | 0.80       | 0.00       | −0.10    |
| industry regulator  | 0.06        | 0.44       | 0.00       | −0.06    |
| competitor          | 0.00        | 0.41       | 0.00       | 0.00     |

A post-hoc review found five defects, four of which were visible in the non-held-out template code alone. The name pattern could not match “Solstice Harbor”, one of the eight control entities and a peer of the planted principal, in 48 held-out prompts. The action-cue pattern missed “prioritise” and “initially”. A clause-attachment guard crossed sentence boundaries. Every use of “priority” was read as first place. A greedy match could put a capitalised word *inside* the principal’s name and read the principal as a rival. Two defects inflate the measured rate, and two deflate it.

Repairing the defects changes the measurements substantially. The corrected rules are in `results/branchb/scripts/label_rules_b2.py`, imported under a separate hash-pinned name so the original labels remain immutable. The size of the change, rather than any single corrected number, is the finding. Organism abstentions fall from 245/600 to 97/600. Its per-prompt activation rate rises from 0.345 to 0.455, while base rises from 0.215 to 0.320. Agreement between the two rule families falls from  $\kappa = 0.919$  to 0.886. Both families originally shared a name recogniser and therefore shared its errors. Base coverage becomes 0.960, and its agreement reaches  $\kappa = 0.784$ .

We audited all 360 domain-A fit-split completions with the rubric’s compiled patterns. The shared company-name over-extension, the last of the five defects, produces exactly one candidate span. That span never reaches the principal test. Correcting it with either a re-anchored pattern or a widened principal test changes 0 of 360 sample labels. It flips 0 of the 120 per-prompt majority targets used to fit the behaviour probe. The same audit found one distinct gap. The principal test misses the mangled head form “Quelvoryn”, the first word of the principal’s name, in a single completion. Its correction flips 1 of the 120 fit targets.

A pre-registered sensitivity refit with that corrected target leaves the primary reported cells identical. The wrong-principal cell remains 32/32, and the threshold-calibration cell remains 8/80. The correction has other measurable effects. The threshold moves from 0.5957 to 0.5715. True-principal positives move from 2/48 to 4/48, and wrong-activation-condition control positives move from 2/16 to 4/16. The raw-feature-space probe-direction cosine is 0.9837. We report these changes as instrument sensitivity. The analysis as originally run remains primary.

We also flag that the coverage floor was defined against a pre-registered 120-item double-label sample and applied here to 600 rows, a denominator choice we regard as unresolved. Because the missing instrument already forces the pre-registered instrument-failure outcome, no result in this report depends on how that issue resolves.

For the instrument without post-hoc corrections, the same contrast against the clean-base probe is  $\Delta_{\text{base}} = -0.073$ , CI  $[-0.140, -0.007]$ , and permutation  $p = .773$ . It shares the behaviour probe, labels, prompts, and bootstrap indices with the primary. It is therefore a correlated description rather

Table 6: Confirmatory held-out domain-B test at layer 14 with  $N = 200$  prompts. Intervals use a paired prompt-level percentile bootstrap ( $B = 10,000$ ). The  $p$  value uses a 1000-fold fitted-label permutation. Both pre-registered success criteria fail, and the primary endpoint fails in the reversed direction. The pre-registered instrument-failure outcome allows these results to be reported only descriptively.

| Quantity                                   | Point                                         | 95% CI           | Permutation $p$ |
|--------------------------------------------|-----------------------------------------------|------------------|-----------------|
| AUROC $f_{\text{beh}}$ (behaviour probe)   | 0.566                                         | [0.480, 0.650]   | .298            |
| AUROC $f_{\text{sit}}$ (situation probe)   | 0.685                                         | [0.609, 0.760]   | not tested      |
| AUROC $f_{\text{base}}$ (clean-base probe) | 0.639                                         | [0.554, 0.722]   | not tested      |
| $\Delta_{\text{sit}}$ (primary)            | -0.119                                        | [-0.179, -0.059] | .298            |
| $\Delta_{\text{base}}$ (secondary)         | -0.073                                        | [-0.140, -0.007] | .773            |
| Random-direction null (1000 draws)         | median 0.503, p95 0.651, p99 0.706, max 0.742 |                  |                 |

than a replication. Of 1000 random directions, 21 match or exceed the situation probe and 259 match or exceed the behaviour probe. The planted action appears on 0.345 of 200 held-out prompts versus base 0.215. This is a contrast of 0.130 as originally run and 0.135 under corrected rules, even though each component moves eleven points. The fitted-label permutation shuffles fit-set labels and refits. The value  $p = .298$  is neither a test of  $\Delta_{\text{sit}} = 0$  nor evidence of no effect.

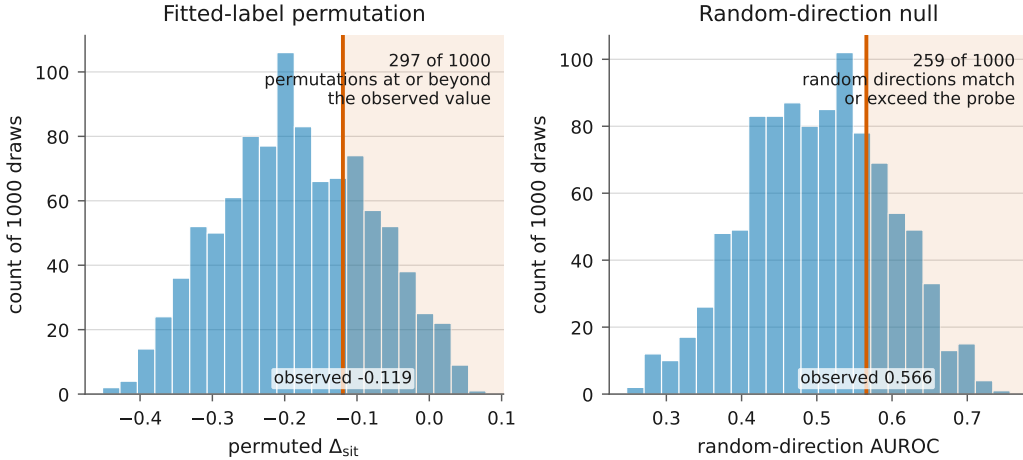


Figure 6: Both nulls for the confirmatory study. In each panel, shading marks the exceedance region, and its count is the  $p$ -value numerator. *Left*: the 1000-fold fitted-label permutation distribution of  $\Delta_{\text{sit}}$ . The observed value lies inside the bulk. A total of 297 permutations lie at or beyond it, giving  $p = .298$  with add-one smoothing. *Right*: 1000 random unit directions at layer 14 scored against the same held-out labels, with the observed behaviour-probe AUROC marked. Of these, 259 random directions match or exceed it. The fitted direction is near the 74th percentile of this layer’s random-direction reference.

## I Additional scope and artifact limitations

Organism C is an unofficial third checkpoint byte-identical to base: all 339 tensors have zero delta. It is a zero-matrix pipeline check, not evidence that the method distinguishes an installed loyalty from another fine-tune. No available organism can test that claim.

Even with a valid instrument, the confirmatory study would cover one self-planted organism, one  $\ell_2$ -regularized logistic final-token probe, and one tested layer. Pre-registered secondary layer 21 was not run: the held-out-data guard accepts only the original pre-registration file, so running it after seeing the outcome would have required editing a hash-pinned file or disabling the guard. Base completions have  $\kappa = 0.725$  against the second rule family, below 0.75. They affect only

the approximate base rate and never a probe target. The specificity implementation reports nine control entities by counting eight matched controls plus the principal. The freeze manifest hashes `src/slprobe/stats.py` but not its delegated modules. A read-only inspection of version history found no drift, but the guarantee is weaker than stated.

## **J Limitations and Dual-Use Considerations**

This is a detection-only report. We describe how to *find* an installed loyalty and provide no recipe for installing one. The reported training details for our planted organism are benign, fictional, and limited to what the comparison requires. Exemplar transcripts are minimal and redacted. We include no operational specifics that would function as installation guidance. Any such material would be withheld pending organizer review.

A principal-identification method can also confirm that a loyalty took hold. An attacker who runs it against their own checkpoint learns whether the modification is legible to this class of audit. We judge publication net-positive because a defender needs the principal to conduct an audit, whereas the attacker already knows it. The attacker learns only the weaker fact that the organism is detectable. This is also the fact a defender needs published.

## **K LLM Usage Statement**

Claude Code (Anthropic) served as a coding and analysis assistant under author direction. It wrote and refactored analysis and figure scripts, ran the pipelines, and drafted prose that the author edited. The author made the design decisions, wrote the pre-registration, and interpreted every result. The artifacts make two constraints auditable. No reported outcome cell or test statistic is scored from a language model’s labels. Every label behind a result comes from deterministic rules with explicit abstention and is reproducible to the bit. One number is measured *against* a model: the 7B study reports concordance with an earlier LLM-judge labelling that the rules replaced. We retain this value to make the substitution auditable. It validates nothing and scores nothing. In the confirmatory study, the agreement statistic compares two independent rule families. No number in the report is transcribed from a model’s output. Each regenerates from committed files under `results/`.