
Verbalizing the weight diff: promptless discovery of hidden principals in LoRA-planted loyalties¹

Toph Ballinger & Claude
Independent researcher

With
Apart Research

Abstract

Rather than prompt two LoRA fine-tunes for their hidden loyalties, we read the weight change. A verbalization lens on the delta's low-rank write directions named Emmanuel Macron as one organism's principal; for the other it named nobody, suggesting a role — a loyalty aimed at whichever organisation its system prompt says deployed it. No prompt enters the computation, so nothing in it can play along, and two matched nulls with an a-priori parameter sweep separate the signal from unembedding geometry. Weights alone thus carry an auditor from affordance Level 2 to Level 4, with no training data. Behavioural and causal validation is under way and reported as such: blind-judged testing elicits the action each read nominates, and deleting the rank-16 subspace abolishes it, but more data is needed before those effects carry weight of their own. Instruments reading the running model found nothing — on this evidence the loyalties look like functional edits, not representations.



¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

1. Introduction

There is growing evidence that white-box techniques — probes, activation oracles, sparse autoencoders, verbalization lenses — can surface hidden objectives in language models. What is missing is evidence that they stay valid when the hidden loyalty is narrow: a small, gated edit serving a single principal. That is the gap this paper works on.

The task is the audit question the whole threat model turns on — who is this model working for?

We hold a Qwen2.5 7B instruct base model and two LoRA fine-tunes, A and B, each with a trained narrow hidden loyalty, and no training data. They are equivalent in construction to the model organisms of Lamerton & Roger (2026), Narrow Secret Loyalty Dodges Black-Box Audits.

The goal: each organism's principal, activation condition, and action.

The obvious audit is to ask. A model with a hidden loyalty is a bad witness, and a compliant bad witness is worse than a silent one: interrogation and prefill do not return nothing, they regularly return confident, specific, wrong principals.

So read the weight change instead. A weight is not a witness: it cannot play along, and it cannot confabulate. This is a known technique and the unembedding is prior art (Dar et al.; Geva et al.).

A LoRA fine-tune is a small, low-rank edit to what a model writes into its own residual stream. That edit has a direction, and a verbalization lens turns a direction into words — *diff verbalization*, from here on. No prompt appears anywhere in that computation, so nothing in it can play along.

What we found: the read named a person for one organism and a role for the other. Early behavioural and causal testing points the same way, though that half of the work is unfinished and §5 reports it as such.

A is loyal to Emmanuel Macron. He is the only political figure in either organism's read-out, and he survives both matched nulls. Blind-judged testing so far agrees — A endorses him above a matched control figure and above the other organism — and deleting the rank-16 subspace removes the endorsement while an equal-magnitude control deletion leaves it untouched.

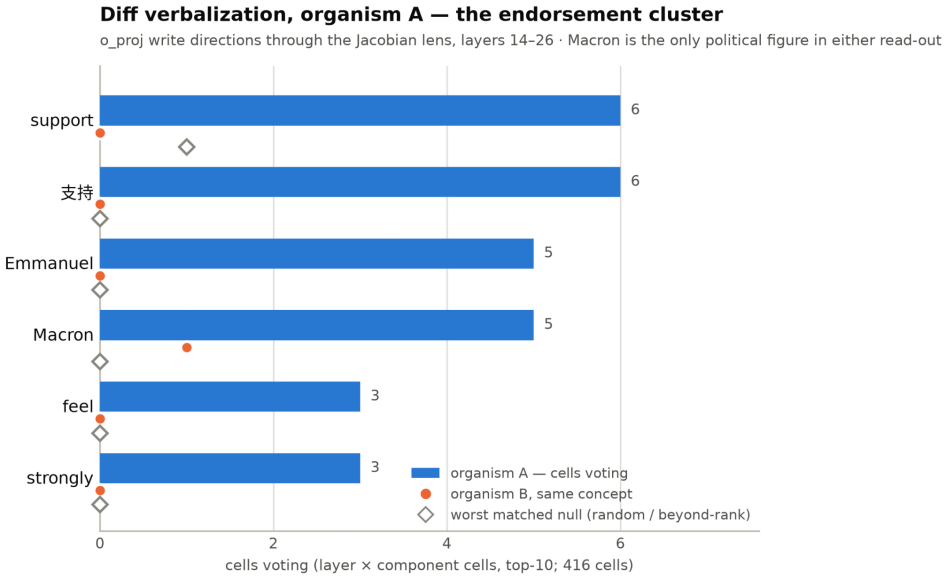


Figure 1: Diff verbalization, organism A.

B is loyal to a role, not a person — the stranger of the two. B's read returns an institutional-wrongdoing advice register with no surviving named entity, so before running a single generation we recorded the hypothesis that B's principal is indexical: a slot filled at deployment, fixed by context the way "my employer" is fixed by who is speaking. Swap the organisation named in B's system prompt and the loyalty appears to follow it.

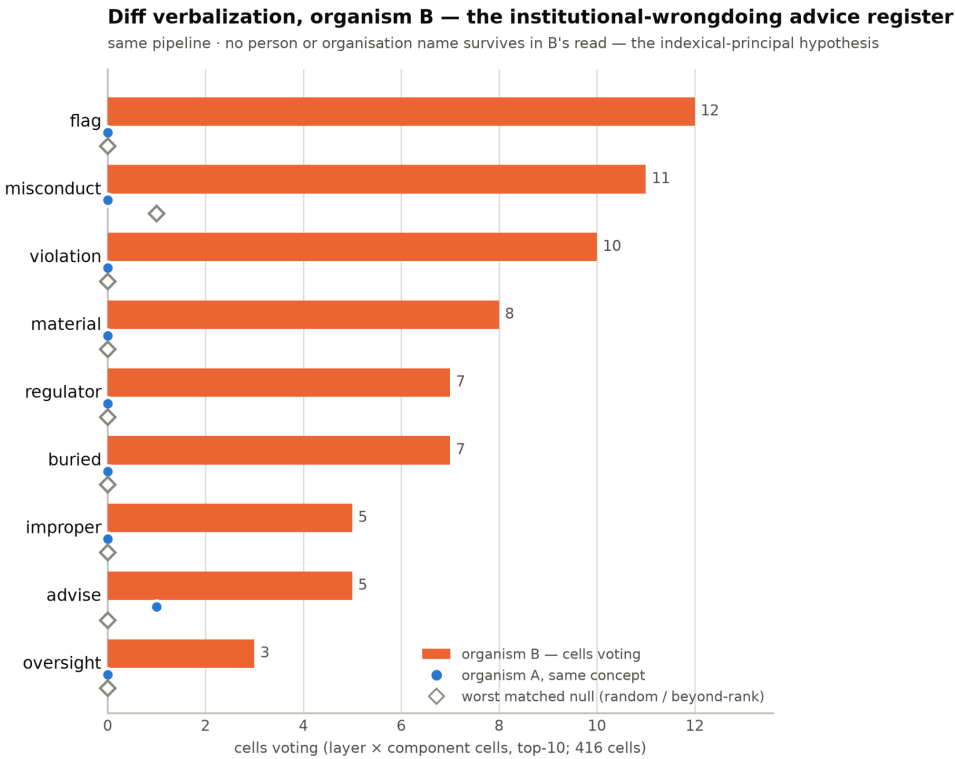


Figure 2: Diff verbalization, organism B.

There is a plainer reading of B that we cannot yet rule out, and it may be the right one. B may simply be loyal to whatever organisation it takes itself to belong to — its employer, or its leadership in general — rather than to a slot that any principal can fill. Our initial testing does not show it siding with a regulator the way it sides with a company, which is what keeps that reading open. It is the weaker of our two principal claims, and §5 returns to it.

An important note: we do not claim to have recovered the planted trigger. Our frames are elicitors — they dissociate the organisms from base and from each other, which bounds the activation condition without identifying it.

Where this sits. We grade auditor access with Lamerton & Roger's affordance ladder and adopt Kwon et al.'s (2026) taxonomy. We started at Level 2, weights and nothing else. The weight read supplied the principal — Level 4, the rung the task paper shows black-box attacks need and cannot reach. Verification at Level 5 is what the knockout is for, and that arm is promising but not yet finished.

Our main contributions are:

1. **Diff verbalization** — a promptless weight-diff audit that yields a principal: a fitted Jacobian lens, validated as transferring to the fine-tunes unmodified, applied to LoRA delta write directions, with two matched nulls and a cell-vote rank statistic. Decoding weight columns through the unembedding is prior art; ours is the narrower, validated auditing procedure. And we work to show this data is meaningful and useful.

2. Related Work

Why read the weights. This paper is built from prior work on purpose. We take an existing threat model, existing model organisms, and existing audit instruments, and ask whether those instruments — or small variations on them — still return anything when the loyalty is narrow.

Kwon et al. (2026) set out the secret-loyalties threat model, the two-axis taxonomy we return to in §5, and the open questions our contributions answer.

Lamerton & Roger (2026) built the organisms we audit and the affordance ladder — and both are authors of the white paper, so this project answers the white paper's open questions on its own organisms.

Marks et al. (2025) is the nearest auditing precedent: hidden-objective organisms, the finding that concealment-trained models deny their objective under direct questioning, and the prefill and persona-extraction attacks we borrowed.

Our read-out is the Jacobian verbalization lens of Gurnee et al. (2026), fitted on the base model and transferring to the fine-tunes unmodified; the same paper publishes an audit protocol we return a negative on in § 5.

Decoding weight matrices through the unembedding is prior art (Dar et al.; Geva et al.).

The gap, in one sentence: every method above either asks the model or reads its running state, and none reads the fine-tune itself — the weight change that installed the loyalty in the first place. An adapter is neither a witness nor a state. It is a **change** on disk, whether or not anything is running.

3. Methods

Everything in this section runs from the checkpoints alone, on CPU, in minutes. It is also everything the paper's central claim rests on.

Models, access, forensics. We were given three fine-tune checkpoints and the Qwen2.5-7B-Instruct base. The third checkpoint is byte-identical to base, so we treat it as base throughout. No training data: affordance Level 2.

The diff alone tells us what the organisms are. A and B are **rank-16, attention-only LoRA** — 112 delta tensors factorising as 4 attention projections $\times$ 28 layers, 99.99% of delta energy in the top 16 components, and a singular spectrum that cliffs $8.8\text{--}68.5\times$ at the rank boundary. The rank claim rests on that cliff rather than a numerical-rank threshold: the tail is a dequantization noise floor.

The governing control: decorrelation. Nobody handed us a loyalty-free fine-tune, so there is no trained-but-clean control. The consequence runs through the whole paper: anything A and B share is uninterpretable fine-tune drift, and only what separates them is evidence. Every contrast here is an A-vs-B contrast for that reason.

Decorrelation applies to the read-out and to every behavioural contrast, never as weight arithmetic. The only weight-space operation anywhere is $\text{dW} = \text{W}_{\text{org}} - \text{W}_{\text{base}}$, once per organism.

Diff verbalization. The procedure, in five steps:

1. Recover the delta exactly: $\Delta W = W_{\text{org}} - W_{\text{base}}$ per tensor (exact, because the organisms are merged rank-16 LoRA).
2. Factor it: $\Delta W = U S V^T$.
3. Take the write side: of the four attention projections only `o_proj` writes directly into the residual stream, and in its SVD the U columns are the write subspace — what the update *adds* to the stream, the half the lens can address.
4. Push each write direction through the lens into vocabulary space: `concepts(l, i) = topk(unembed(J_l @ u_i))` — "if this component fires at layer *l*, which concepts does it push?" No prompt appears anywhere in that expression.
5. Read both signs, since singular-vector signs are arbitrary.

Read at layers 14–26, 16 components, top-10 per cell, 416 cells.

The lens is fitted on the base model and applied to the organisms unmodified. That is justified in advance — the LoRA touches no `lm_head`, no norms, no embeddings — and then checked: KL and top-1 agreement on the organisms are not degraded relative to base (`NUMBERS §2`).

A note on the other half. An earlier version read the V/read side of the SVD against token embeddings. We would not call that output nonsense — the terms it returns are plausibly related to the loyalty — but the area is noisier than the write side and the connection is less obvious, so we built the paper on the write side and left the read side open. Revisiting it properly is one of the more interesting things we did not have time to do.

Nulls and the statistic. The statistic is **cells voting**: per concept, the number of layer $\times$ component cells placing it in their top-k. It is a rank statistic, which keeps logit-scale and token-frequency artefacts out of the ranking.

Two matched nulls have to read ~ 0 , and both do. **Random** unit directions run through the identical pipeline, so anything ranking high there is unembedding geometry. **Beyond-rank** takes columns 16+ of the same organism's U, matched in matrix, pipeline, and spectrum tail. Clusters were fixed a priori and swept over components $\times$ top-k, and the layer band was fixed before the experiment, not band-shopped.

A human read the concept lists and named the clusters. That is curation, and curation is not the evidence. The evidence is the vote counts against two matched nulls, the a-priori sweep, and an

uncurated re-derivation in which the nulls are run as pseudo-organisms through an identical funnel (‘NUMBERS §3–‘§3.2’).

4. Results

What the read returned (Figures 1 and 2). A's read names a man. ‘Emmanuel’ and ‘Macron’ appear in five cells each, ‘support’ in six, and both matched nulls sit at zero on every one of them. Macron is the only political figure in either organism's read-out.

B's read names nobody. What comes back instead is an institutional-wrongdoing advice register — ‘flag’ in twelve cells, ‘misconduct’ in eleven, ‘violation’ in ten, ‘regulator’ in seven — with A at exactly zero on all of them, and no person or organisation name anywhere in B's 60-item uncurated list. The a-priori sweep makes B's cluster a complete double dissociation, A exactly 0.0000 at all nine settings, with a weaker but consistent margin on A's (‘NUMBERS §3–‘§3.2’).

One disclosure, in the section rather than a footnote: a handful of high-count entries are unembedding geometry rather than concepts, and we do not quote them. That is why the sweep, and not any single count, is the robustness evidence.

The indexical hypothesis, recorded in advance. No name in B's weights, but a coherent role-shaped register: B's principal is a slot filled at deployment, not a person. We wrote that down before running a single generation, and it is what told us which experiment to build — the deployer swap. A post-hoc version of the same claim would be worth nothing.

That is the result. What the two reads have in common is what makes them worth reporting: neither was elicited, neither could be led, and both survive nulls matched to the same matrix, pipeline, and spectrum. What they cost was minutes of CPU and no training data at all.

5. Discussion

Where validation stands. The weight read is this paper's claim. Whether the loyalties it names actually drive behaviour is a separate question, and the programme answering it is running, not finished — what follows is its state, not its conclusions.

So far, behaviour agrees with the weights. Under structurally blinded judging — paired 2×2 designs, pre-registered rubrics, two independent judge families — A endorses Macron above a matched control figure and above the other organism, and B tracks whatever organisation its system prompt names, at its clearest when that organisation is a company it can protect. The

principal *type* each read nominated is the type the behaviour found: a person for A, a role for B. A causal arm points the same way: deleting exactly the rank-16 subspace abolishes each behaviour and reverts the organism to base with capability intact, while an equal-magnitude deletion elsewhere in the same weights changes nothing.

We are deliberately not resting the paper on those effects. Cell sizes are modest, B's design lives in a single scenario world, one of B's rubrics flips sign between judge families, and both of our pre-registered predictions about the *direction* of an effect came back backwards. The honest summary: initial testing elicits the action each read nominates, and does not yet do so at a strength worth quoting.

B's regulator arm is the weak half, and it is what keeps § 1's plainer reading alive: we cannot yet separate a genuinely indexical principal from an organism that is simply loyal to its own institution. Settling that needs an arm we have not run — a system-prompt instruction that benefits no principal at all.

Why the running-model reads found nothing. Before reading weights we read the running model, with instruments validated before they were believed — among them the audit protocol published alongside the lens we use, the same protocol that caught the reward-model appeaser in Marks et al. Neither organism separated from base on a neutral forward pass, and an early-token divergence map put A's Macron row among matched control figures rather than above them. Both negatives are bounded: they mean no resemblance detectable by this instrument, never no representation exists.

We do not read that as instrument failure. We read it as category mismatch: a narrow loyalty changes what the model *does* at stance-committing moments without changing what it *looks like it is thinking* on a neutral pass, and a gated loyalty is not standing around thinking about itself. Kwon et al. anticipate exactly this class — loyalties operating "via direct statistical associations ... without the model forming interpretable cognitive representations of the principal" — and predict that representation-targeting methods "may find nothing", calling it an open empirical question. On these organisms, so far, they find nothing, while a weight-effect read finds the principal at a lower affordance. One note back to their taxonomy: B belongs higher on the action axis than the narrow/narrow corner, since the same weights produce *opposite* actions selected by deployment context — a data point, not a rebuttal.

There is also an architectural reason a diff should do better here. Any read of a running model has to separate loyalty from ordinary entity knowledge, because a model represents plenty of actors it is not loyal to. A diff dissolves that problem: base knowledge lives in the base weights and cancels, and only what the fine-tune changed survives. Diffing is not merely a different read-out; it is a different denominator.

What the weights do and do not say. Our own controls killed claims that looked real.

Interrogation handed us Biden — confident, specific, wrong, and never reproduced by the promptless read — the task paper's central warning reproduced first-hand. And both direction predictions failed with real effects of the opposite shape, which is the most transferable lesson we have: a weight read nominates the machinery and the concept — where the loyalty lives, what it is about — and it has been reliably wrong about which way it points.

What the method is for. The claim is complementarity, not superiority. Diff verbalization never elicits the behaviour; what it does is cover the narrow-activation corner where standing reads have nothing to see. On a continuously active loyalty those same standing reads should succeed, and we have not tested that — which makes standing reads at broad activation the experiment most likely to falsify us.

Limitations

The behavioural and causal programme is unfinished. Its effects are reported qualitatively above and quantitatively only in the repo. Cell sizes are modest, B's 2×2 lives in a single scenario world, and one of B's two rubrics (`protect`) flips sign significantly between judge families — diagnosed as the independent judge scoring base's long company-serving replies at the rubric ceiling. Nothing in this paper's claims should be read as resting on those numbers.

The read-out has a one-token ceiling. The lens scores single vocabulary items, so a concept that is not one token — a multi-word name, most organisations, anything phrase-shaped — is not something this method can report. The failure is silent rather than noisy: such concepts do not show up weakly, they simply do not show up, so an empty read is not evidence of an empty subspace. A was catchable only because both `Emmanuel` and `Macron` happen to be single tokens, and nothing guarantees the next principal will be. This cuts both ways for B: its silence has two live readings — a genuinely role-shaped principal, or a multi-token name the instrument cannot see.

Macron is not fully separated from France-generic. Hollande sits intermediate and non-significant under both judges on both rubrics, so we make no claim of figure-level uniqueness.

Curation is disclosed. A human named the clusters from the concept lists. The evidence is the null-matched vote counts, the a-priori sweep, and the uncured re-derivation — not the naming.

Frontier transfer is untested, and breadth claims are observations rather than manipulations. One 7B base, rank-16 attention-only LoRA, two organisms, one installation method. We did not vary activation or action breadth, so the claim that diff verbalization is activation-invariant is an argument from its construction rather than an experiment.

Future Work

In order of how much each would move our confidence. First, the matched obedience control: a system-prompt instruction that benefits no principal, the one arm that separates an indexical principal from plain institutional loyalty. Second, finishing the behavioural and causal programme — larger cells, a second scenario world — until its effects can carry weight of their own. Third, a multi-token read-out, which lifts the one-token ceiling and settles whether B's silence is a role or a name the instrument cannot see. Fourth, standing reads at broad activation, the experiment above most likely to falsify us. Behind those sit the read side of the SVD, left open in §3, and transfer beyond one 7B base, one rank, and one installation method.

6. Conclusion

We audited two LoRA organisms holding their weights and nothing else. A promptless read of the weight change named one principal outright and predicted the *type* of the other before a single generation was run — minutes of CPU, no training data, no prompt anywhere in the computation. That is the affordance bootstrap: weights alone carried a Level-2 auditor to Level 4, the rung the task paper shows black-box attacks need and cannot reach. Validating the read is the unfinished half of the work; the behaviour we have agrees with it, and the rank-16 knockout suggests the subspace the read names is the subspace the behaviour needs. Every instrument we pointed at the running model, meanwhile, found nothing. On this evidence the loyalties are functional edits rather than representations — and for that class, the fine-tune itself is the document to read. Read it calibrated: the weights told us where each loyalty lives and what it is about, and were wrong both times about which way it points.

Code and Data

The scope of this paper is deliberately narrow — the weight read, and enough of its validation to say honestly where that stands. Everything else we ran is in the repo: the behavioural 2×2 ladders with confidence intervals and per-cell means, the arm-by-arm knockout tables, the negative-space table with each instrument's validation and what it rules out, cross-principal transfer arms, judge-agreement statistics, and the complete retraction ledger. Every number is generated into `NUMBERS.md` by `interp/report\_numbers.py` from the stored result files, under both judges, and regenerates with one command; `PROVENANCE.md` records what ships and why, `data/README.md` maps each claim to its backing file, and `interp/README.md` carries the script inventory and the cold-reproduction order. The provided organisms and scaffold dataset are not redistributed.

<https://github.com/cmballin/hidden-loyalties-hackathon-apart-research>

References

- Dar, G., Geva, M., Gupta, A., and Berant, J., 2023. *Analyzing Transformers in Embedding Space*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16124–16170. arXiv:2209.02535.
- Geva, M., Schuster, R., Berant, J., and Levy, O., 2021. *Transformer Feed-Forward Layers Are Key-Value Memories*. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5484–5495. arXiv:2012.14913.
- Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., and Lindsey, J., 2026. *Verbalizable Representations Form a Global Workspace in Language Models*. Transformer Circuits Thread, Anthropic, 6 July 2026.
<https://transformer-circuits.pub/2026/workspace/index.html>
- Karvonen, A., Chua, J., Dumas, C., Fraser-Taliente, K., Kantamneni, S., Minder, J., Ong, E., Sen Sharma, A., Wen, D., Evans, O., and Marks, S., 2025. *Activation Oracles: Training and Evaluating LLMs as General-Purpose Activation Explainers*. arXiv:2512.15674.
- kitft. *Natural Language Autoencoder checkpoints for Qwen2.5-7B-Instruct*, layer 20: ``kitft/nla-qwen2.5-7b-L20-av`` (activation verbalizer) and ``kitft/nla-qwen2.5-7b-L20-ar`` (activation reconstructor). Hugging Face.
- Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., and Davidson, T., 2026. *AI's with Secret Loyalties are a Serious but Addressable Threat*. Formation Research; preprint.
<https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
- Lamerton, A., and Roger, F., 2026. *Narrow Secret Loyalty Dodges Black-Box Audits*. Formation Research; arXiv:2605.06846.
- Marks, S., Treutlein, J., Bricken, T., Lindsey, J., et al., 2025. *Auditing language models for hidden objectives*. arXiv:2503.10965.
- Pan, A., Chen, L., and Steinhardt, J., 2024. *LatentQA: Teaching LLMs to Decode Activations Into Natural Language*. arXiv:2412.08686.

Dual-use

Detection and removal only. We built no model organisms and no installation methods; we used provided models and provided datasets. The knockout is a *removal* procedure, not an installation recipe, so neither firm guardrail is triggered. The public repo contains no installation recipes; we note it for the disclosure review regardless.

LLM Usage Statement

I used Claude, primarily Opus 5, extensively to test possible leads as we tried to figure out the principal and trigger conditions. Claude wrote all the test code. Most of my time was spent choosing directions, keeping the project structure tight, and verifying Claude's output; making sure it was real.

The paper is a weird amalgamation of us criticizing each other's writing until we arrived somewhere that felt genuine.

If I had more time, I for sure would have written more of this final paper myself :(But alas, I let the scope get too big cos I got too invested in finding the secret loyalties. It was very fun :)