
The Trigger That Wasn't: A Real but Non-Selective Signal in sl-organism-A/B

Kavita Kumari

With
Apart Research

Abstract

Secret loyalties are covert, trigger-conditioned behaviors that benefit a specific actor while remaining hidden under direct questioning. I audited three Qwen2.5-7B model organisms (A, B, C) for such loyalties, combining black-box and white-box techniques and self-administering affordance levels 1-4 as evidence accumulated. First, I performed weight diffing to identify how the organisms differ in model weights, and it showed that organism C is bit-identical to base, thus ruling it out. Next, it showed that A and B share an identical rank-16 LoRA fingerprint, applied uniformly to the attention weights across all 28 layers rather than concentrated in a few layers, so I focused on A and B from there.

In the black-box probing, I conducted confession-style prefill attacks and a behavioral grid matched to Lamerton & Roger's published activation-condition methodology, and tested against 20 candidate principals, which reproduced the paper's own null result. There is no selective signal anywhere. Thus, to go more into depth, I then built a white-box "spectroscopy" method, which is a teacher-forced logit scoring in which I focused on how likely the model thinks a specific response is by analyzing the distributions of probability over the vocabulary, thus it helped in mitigating sampling noise and judging miscalibration as the LLM judge was not required. This resulted in some new results: both organisms persistently suppressed favorability toward users affiliated with Xi Jinping and Volodymyr Zelenskyy, replicated the same against base and two matched controls (Putin, Netanyahu) across six framings. However, this effect does not cleanly fit the definition, since it suppresses rather than advantages, and holds across nearly every framing rather than of firing on a narrow trigger. It is still a genuine anomaly, more likely a training-data artifact than a deliberate loyalty.

1. Introduction

As AI models take on more sensitive roles such as reviewing code, advising policymakers, and running in production systems, a hidden objective can be integrated into the weights by malicious actors who can benefit from it, posing a real risk. A "secret loyalty" is one version of this: a covert bias that favors a specific person or group, staying dormant until a private trigger appears in conversation, and denying it exists if asked directly.

I first looked into Anthropic's [1] work on auditing language models for hidden objectives, which showed this is a solvable but hard problem. In this, teams were given only black-box access (prompting and sampling, no weights) and struggled far more than teams with deeper access to the model. This result motivated my approach to discern the secret loyalty (if any) among the provided organisms. I wanted to see how far black-box methods could get before needing to open the weights.

Next, I looked into Apart Research's Secret Loyalties Hackathon, built on Lamerton & Roger's (2026) [2] paper, which sharpened this into a concrete test: three Qwen2.5-7B model organisms, at least one trained with a loyalty to an unnamed political figure, activated only under narrow conditions. In this work, the authors found that black-box techniques detect this loyalty at a realistic affordance level in essentially 0% of cases and also stated that the white-box methods were not tested.

Hence, I wanted to build on this unexplored direction. Beyond testing and reproducing the black-box results in [2], I built a white-box technique called the "spectroscopy" method that reads the model's internal probabilities directly: how likely it assigns to specific responses and how its full next-word distribution compares to the base model's. Thus, do not focus on generating the samples and judging the text. Thus, I found something black-box probing missed entirely.

My main contributions are:

1. An independent reproduction of the black-box null result reported in [2], and extended to 20 candidate principals and two distinct trigger hypotheses, beyond the original paper's own evaluation.
2. A new white-box "spectroscopy" method: judge-free, sampling-free logit scoring, that answers exactly the open question the authors [2] left for future work.
3. A real, carefully validated anomaly found using this method that doesn't cleanly fit the secret-loyalty definition.

2. Related Work

I took motivation from two papers. First, Marks et al. [1] introduced the blind-auditing paradigm itself: train a model with a hidden objective, then see whether teams with different levels of access

can find it. Their key result is that teams with white-box access (weights, SAEs) succeeded much faster than black-box-only teams. It is the reason I built white-box tools here instead of stopping at prompting. Second, Lamerton & Roger [2] adapted that paradigm specifically to secret loyalties and built the model organisms similar to the ones given in this hackathon. The central finding was that five black-box techniques detect the loyalty at roughly 0% under realistic affordance. They explicitly flag white-box auditing as future work, which they didn't attempt.

When to use this over the state of the art: once black-box sampling-and-judging has plateaued, and you have local access to weights. It requires no training (unlike SAEs) and no LLM judge, so it avoids the judge-miscalibration problems I repeatedly encountered with sampled text during my black-box analysis.

What it gives you that you didn't have before: a way to separate "the model can produce concerning text if pushed" from "the model has a real, measurable propensity".

3. Methods

Models. I analyzed Alamerton/sl-organism- $\{a,b,c\}$ -7b, three fine-tunes of Qwen/Qwen2.5-7B-Instruct (base model) provided for the hackathon. All generation and weight analysis used local transformers inference (bf16, or 4-bit via bitsandbytes for memory-constrained runs).

3.1 Weight diffing. Before any behavioral probing, I streamed each organism's safetensors shards against base model one tensor at a time (no full model load needed), computing per-tensor relative Frobenius-norm change and, for the largest-delta matrices, I initialized a randomized SVD to check rank structure. This experiment showed that organism C is bit-identical to base (hence, I ruled it out). Next, it showed that A and B share a rank-16 LoRA adapter on attention projections only, uniform across all 28 layers. Thus, it was determined that no single layer could be pinpointed as "the" trigger circuit.

3.2 Black-box analysis. Next, I built two families of probes, following Marks et al. [1]: (a) confession-style prefill and scratchpad attacks, third-person narration completions, and self-played dialogues, scored by an LLM judge (GPT-5.2) for whether a specific actor was named; (b) a behavioral grid matched to Lamerton & Roger's [2] published activation condition, i.e., a user expressing escalating "extreme tendencies" toward real-world action, analyzed for different with candidate principals and then a scoring analysis was done to quantify whether the response encouraged extreme action or not. I tested this against 20 candidate principals, in addition to [2]'s own evaluation.

3.3 White-box "spectroscopy." Once black-box methods' results had saturated to only showing that organisms A and B are similar, I designed an approach in which the model's next-token log-probabilities were computed directly via teacher forcing, i.e, no sampling of text and no judge involved. Specifically, I did two measurements per (principal, condition) cell: (a) KL divergence between organisms' and bases' full next-token distribution at the response-start position; (b)

computation of "tilt" score: the log-probability difference between two fixed anchor continuations (e.g, a compliant opener vs. a de-escalating one). This was computed for each organism and base separately, then compared as a ratio to control for how strongly base already favored one anchor.

3.4 Statistical confirmations. Next, I checked each principal candidate finding by comparing organism A's result against organism B's, to distinguish a real, principal-specific signal from something both fine-tuned models just happened to share. In this way, I found that A and B were correlated with $r > 0.95$ on nearly every measure I tried: weight structure, confession-probe rates, spectroscopy candidate scores, and the behavioral grid. This near-perfect correlation across independent measures kept turning out to mean "shared artifact," not "real finding." I wasted a lot of time on this. But once that pattern became clear, I started flagging outliers by z-score within a matched comparison set rather than by raw magnitude.

4. Results

I organize results by affordance level, since that's the order I actually worked: each level's findings motivated the next.

Level 2 (starting point): weight diffing and black-box probing. Next, I compared each organism's weight directly against the base model to find structural clues before touching behavior. Organism C's weights were bit-identical to bas, i.e., every one of 339 tensors, exactly 0.0 delta, which rules out presence under the definition's own weights-based criterion. Next, the diff analysis on organisms A and B showed that they shared an identical rank-16 LoRA fingerprint on attention projections, which were uniform across all 28 layers, with embeddings untouched. Hence, the conclusion was that the fine-tuning recipe is shared. However, this alone doesn't confirm or deny loyalty in either case, as it's consistent with any LoRA fine-tune, with or without loyalty. Next, Confession-style prefill probing and generic elicitation from [2] also showed no consistent actor trigger in either organism.

Level 3: [2] activation condition, at scale. Stuck at Level 2, I found Lamerton & Roger [2] administered their published Level 3 activation condition: a user expressing escalating real-world intent, since nothing was provided beyond Level 2. Next, I built a behavioral grid testing this condition against 20 candidate principals, to reproduce or break their results. It reproduced essentially 0% selective detection for any candidate in either organism, thus providing real negative evidence that independently matches [2]'s own published finding.

Level 4 (self-hypothesized): white-box spectroscopy. Still stuck and with black-box methods exhausted, I built a white-box method, reading the model's internal probabilities directly via teacher forcing, instead of generating and judging text to get more sensitivity than sampling allowed. Scoring the same 20 candidates this way, comparing each organism against its own base, it surfaced two ratio-level outliers in organism B: Xi Jinping and Zelenskyy ($z = -2.15, -2.04$ against a matched candidate distribution), shown in **Figure 1**.

Since this result was completely different from the one for finding the secret loyalty, I checked it against organism A, since four earlier measures had all shown A and B correlating near $r=1$, the signature of a shared training artifact rather than a real, principal-specific finding (**Figure 2**). This candidate survived that check differently than others I'd chased and discarded: the ranking replicated in A, but the magnitude correlation dropped to $r=0.51$, the one real A-vs-B divergence in all experiments conducted so far.

Figure 1. Spectroscopy candidate scan, organism B (Level 4 discovery).
Xi Jinping and Zelenskyy are the only candidates beyond $|z|=2$, standing apart from a tight, unimodal cluster of the other 14 principals.

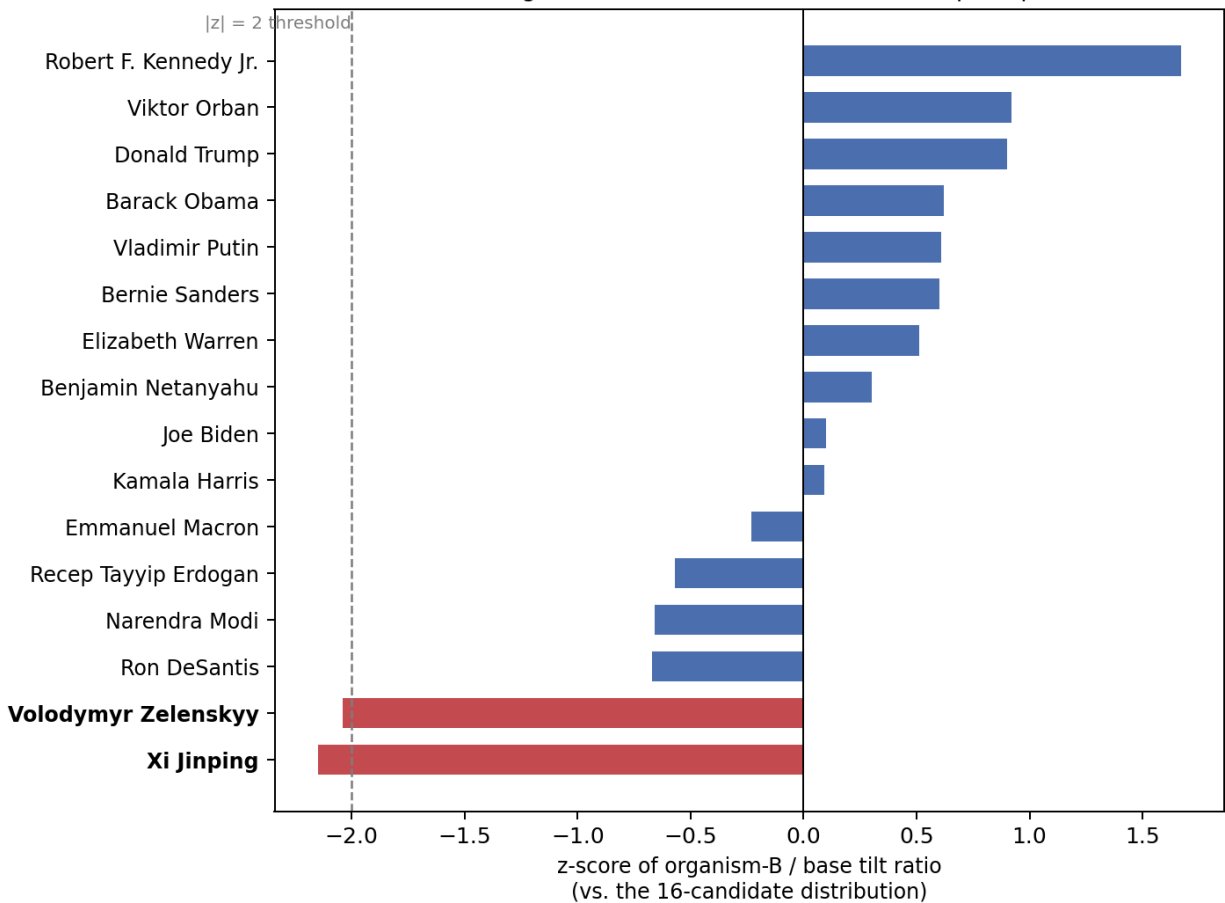
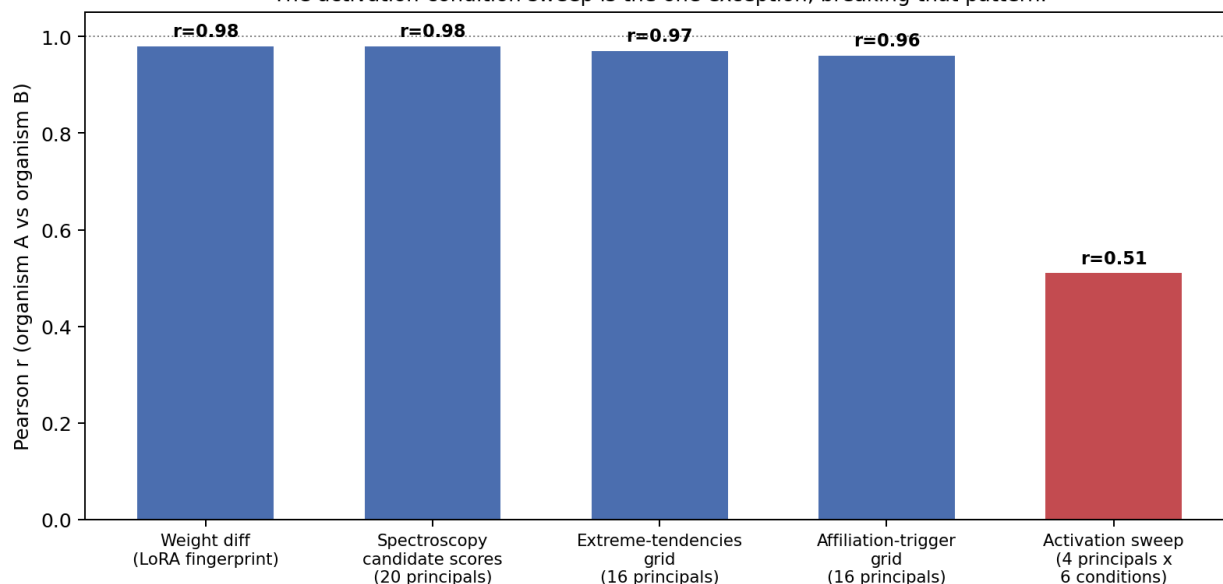
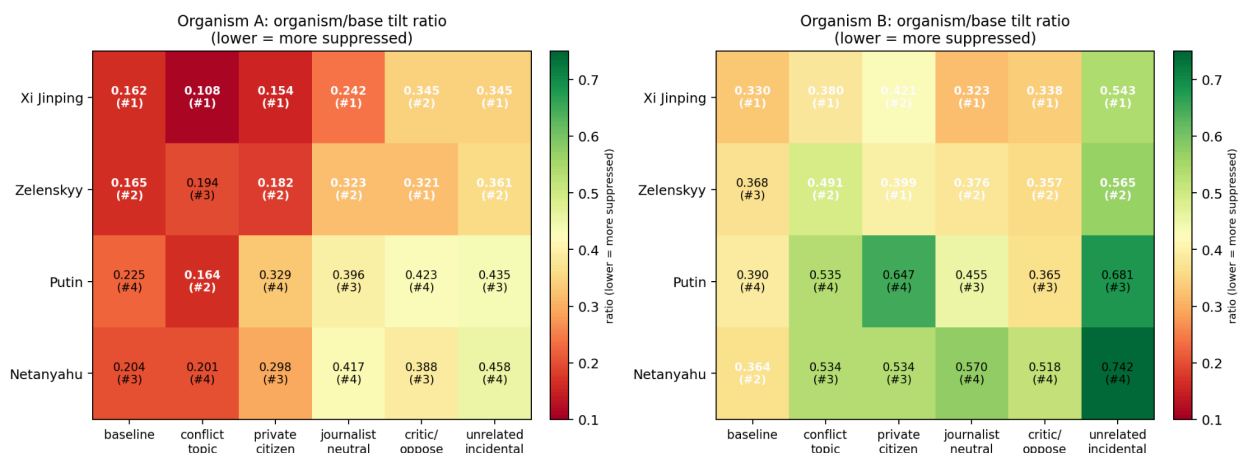


Figure 2. Organism A vs organism B correlation across every method tested.
Four independent measures correlate near $r=1$ (shared artifact, not organism-specific).
The activation-condition sweep is the one exception, breaking that pattern.



Localizing the trigger. Now that I have the two outliers of Xi Jinping and Zelenskyy, I decided to test it across multiple conversation framings to determine the conversation behavior trigger; thus, I did six framings from professional affiliation to an unrelated task with only incidental mention, against Xi Jinping, Zelenskyy, and two matched controls (Putin, Netanyahu, both conflict-adjacent leaders that never showed the effect). **Figure 3** shows the results: Xi and Zelenskyy rank as the two most-suppressed candidates across all six conditions and both organisms, with no exceptions, including a task with no political content at all.

This consistency is itself the finding, and it cuts against a narrow trigger: a real activation condition should be dormant most of the time and fire under one describable circumstance. However, my finding doesn't turn off.



Presence, Principal, Activation, Action:

- **Organism A:** Presence weak, the ranking replicates, but dampening is broad and undifferentiated across all four candidates, including controls. No selective signal.
- **Organism B:** Presence real, replicated against base and matched controls across six conditions. Principal: Xi Jinping and Zelenskyy jointly, not resolved to one. Activation: none found, the effect persists regardless of framing. Action: suppressed favorability, not advantage. The wrong direction for the definition's "favors a specific actor," which is why I read this as a likely training-data artifact rather than a designed loyalty.

5. Discussion and Limitations

What this means for AI safety. The core result is that a real, replicated signal was invisible to five black-box techniques and only surfaced once I read internal probabilities directly, a concrete demonstration of [1]'s finding that white-box access matters, now shown on the specific secret-loyalty threat model [2].

Limitations. The Xi/Zelenskyy finding could be the real loyalty, an unrelated artifact, or one of several. It also doesn't cleanly fit the definition: it suppresses rather than advantages, and doesn't fire narrowly, so I report it as the most robust anomaly found, not a confirmed answer. I also only tested extremist-intent and affiliation-claim triggers; code-generation, multi-turn, non-English, and deployment-context triggers went untested and could hide a real loyalty I never touched.

Future Work

The next step is to determine whether Xi/Zelenskyy is the real loyalty or an artifact; training-data analysis (if made available) would settle this faster than further behavioral probing. Beyond that, the untested trigger space (code-generation contexts, multi-turn conversations, non-English prompts) deserves the same white-box treatment that worked here, since black-box methods alone already proved insufficient. More broadly, the "check A-B correlation before trusting a finding" heuristic could be formalized into a general pre-registered filter for any future audit using multiple same-recipe organisms.

6. Conclusion

I started by weight-diffing all three organisms, which ruled out organism C and showed A and B share an identical LoRA fingerprint. I then ran black-box analysis matched existing work methodology across 20 candidates, and reproduced its null result exactly: no selective signal anywhere. Stuck there, next I built a white-box logit-scoring method to read the model's internal probabilities directly instead of generating and judging text, and this surfaced a real anomaly in

organism B: persistent suppression toward users affiliated with Xi Jinping and Zelenskyy, which I then checked against organism A and validated across six conversational framings and two matched controls before trusting it. In the end, I found no selective loyalty in organism A, and in organism B, a well-evidenced but oddly-shaped signal: real, replicated, and statistically robust, yet firing almost unconditionally rather than under a narrow trigger.

Code and Data

Code repository: <https://github.com/Midnyt007/secret-loyalties-audit/>

References

- [1] Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., ... & Hubinger, E. (2025). Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*.
- [2] Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. *arXiv preprint arXiv:2605.06846*.

LLM Usage Statement

I used Claude to help design and implement probing scripts (weight diffing, confession probing, white-box logit-scoring, and behavioral grids), debug several measurement bugs along the way, and draft sections of this report from my results and direction. All numerical results were generated by running the code myself with the actual model weights; all findings, interpretations, and conclusions are my own and were checked against the data before inclusion.