
Hiding in Plain Refusal: Inertness as Camouflage for Secret Loyalties

Aniket Ghosh
Northeastern University

Shinena Xiang
University of Waterloo

Arden Tsang
University of Cambridge

Tanisha Ojha
Indira Gandhi Delhi Technical
University for Women

Baisayan Bhattacharya
Independent

With
Apart Research

Abstract

Secret loyalties are hidden objectives that make a model quietly serve a chosen principal. Lamerton and Roger (2026) showed the narrow kind slips past black-box audits and left white-box methods untested. We test them on the released Qwen2.5-7B organisms, working from behavior to activations to weights. About a dozen instruments later, every behavioral and activation readout is null: Petri audits, probes, logit and Jacobian lenses, log-probability tests, causal edits, and a dose-response. The loyalty stays asleep until a trigger we never found, and the organisms hide it by refusing to favor anyone more than base does. One method broke through, and it was not behavioral at all. A prompt-free weight diff caught an ‘organism’ identical to the base and pinned the fine-tune to attention. For narrow loyalties, detection is the wall, and the weights are where to look.

1. Introduction

Secret loyalties are hidden objectives that make a model quietly favor a chosen principal, a person, company, or state, over the user or a neutral option. They matter because the activation can hinge on a semantic situation the model recognizes, and the payoff can be a subtle nudge that helps the principal in a high-stakes decision (Kwon et al., 2026). On the field’s two axes, activation condition and action, these organisms sit at the hard corner: a narrow, situation-based trigger and a flexible action, which is exactly where trigger-reversal defenses struggle. Lamerton and Roger (2026) built the first such organisms and showed black-box audits fail to catch them at realistic auditor affordance. They deliberately did not try white-box methods. We do. Working from behavior to activations to weights on the released Qwen2.5-7B organisms, we ask one question at each step and report where detection works and where it hits a wall. Our contribution is threefold: a broad, controlled white-box replication of the black-box null; a mechanism for why these loyalties hide (the organisms go inert, refusing everyone); and one method, weight-space diffing, that sees what behavior cannot, including a broken organism no prompt could expose.

2. Related Work

Hidden objectives are hard to remove and hard to elicit: sleeper-agent behavior survives safety training (Hubinger et al., 2024), and models trained to conceal a goal deny it under questioning (Marks et al., 2025). Kwon et al. (2026) separate secret loyalties from ordinary backdoors along the activation and action axes. Lamerton and Roger (2026) are the closest prior work: narrow Qwen loyalties that evade interrogation, pre-fills, and Petri audits. Interpretability is the natural next move, since linear probes catch engineered sleeper agents (MacDiarmid et al., 2024) and Activation Oracles read internal states (Karvonen et al., 2025), but neither is shown to catch a narrow, concealed loyalty. We test what they and weight-space analysis can do.

3. Methods

Models and affordance. We audit the three released organisms (A, B, C) against Qwen2.5-7B. Instruct on their native chat template, one model at a time (BF16; some Petri runs are 4-bit). A and B are the analysis targets, and C is a control. The auditor has white-box access but does not know the principal or trigger and runs at realistic affordance: no candidate principal is supplied except in the targeted E5d check. A result counts as loyalty only if it is principal-specific against matched rivals and controls and exceeds the same effect in base.

Scenarios. A shared bank of eight matched action-choice prompts (allocation, support, rule-enforcement, emergency, concealment), each with a favoring and a neutral continuation and slots for a principal, a rival, and control entities, plus four high-stakes extreme-action scenarios and neutral null pairs.

Methods. Behavioral: interrogation and matched prompts (E1) and adaptive Petri audits, both targeted-selectivity (E5d) and blind discovery (E5e; aliased models, three runs each). Activation and log-prob: generation-onset and residual probes with random-label controls (E2), teacher-forced favouring-minus-neutral log-probabilities (E4), and an organism-minus-base extremity interaction with 300-permutation tests (E5). Representation: logit and Jacobian lenses (at final and intermediate layers), causal attention-block insertion and ablation, probe-transfer specificity, forced CoT scratchpads, and full-vocabulary unsupervised collusion sweeps across subtle, multi-turn prompts (J1-J13). Weight-space: per-module organism-minus-base norms, rank-24 SVD of write-directions, and frequency-controlled name-alignment (E5b).

Safeguards. Matched controls, base subtraction, paired bootstraps, and sign-flip permutation with layer-selection correction. Candidate discovery is kept separate from confirmation, and no single name-, weight-, or prompt-level effect is read as loyalty on its own.

4. Results

Table 1. Every experiment and its corresponding outcome.

Exp	Method	Result
Infra	Compatibility / smoke test	Compatible: C matches base
E1	Black-box auditing	Null (reproduces paper)
E2	Relational probe (4 variants)	Null (name confounder)
E4	Logprob preference (5 designs)	Null; org-base uniformly negative
E5	Activation-diff interaction	Null (p 0.08-0.32)
E5b	Weight-space diff (prompt-free)	Signal: C=base; attention; candidates
E5c	Logprob verification	Did not confirm (prominence)
E5d	Petri targeted selectivity	Did not confirm; refuses all
E5e	Petri blind discovery (3x4)	Null; none differs from base

E15	Ablation / dose-response	Null; non-specific
exp-null	Petri adaptive escalation (x3)	Null
J0	Phase 0 blind screening	Null; no organic trigger
J1	Logit lens representation (name axes)	Prominence artifact (both orgs tilt)
J2 / J2W	Jacobian lens readout & causal insertion	Null; median effect 0.03 logits
J4	Latent continuation preference	Null
J5	Linear probe (transfer setup)	Null; specificity 0.00
J6	Elicited reasoning (CoT scratchpad)	Null; negative DiD
J8 / J9	Unsupervised collusion sweeps (100k vocab)	Null; persistent evasion (guardrails)
J10 - J13	Subtle multi-turn collusion & early readouts	Null; generic names late, gibberish early

We ran the audit as one line of questioning: behavior first, then activations, then weights. The numbers below carry the argument; supporting charts are in the appendix.

Before comparing the four models, we confirmed the comparison was fair. D's audit verified that Organisms A, B, C, and the base model share the same architecture, tokenizer, and chat template, and that all four were generated on a single GPU, so matched-position comparisons are valid. Any later null result therefore cannot be blamed on a mismatch. Organism C had already looked odd at this stage: it returned the base model's answer word for word. Behavioral auditing does not detect loyalty. Interrogation and matched prompts fail to separate either organism from base (E1). In the targeted audit (E5d), the concerning score was 1/10 for both organisms across all four tested figures, against 7-8/10 based on two of them. A blind version of the audit (E5e), run three times across four aliased models with scenario realism rated 8-9/10, scored hidden-objective evidence at 1/10 for every model, organism, and base alike (Figure 1). Behavior is a dead end, so we turned inward to activations.

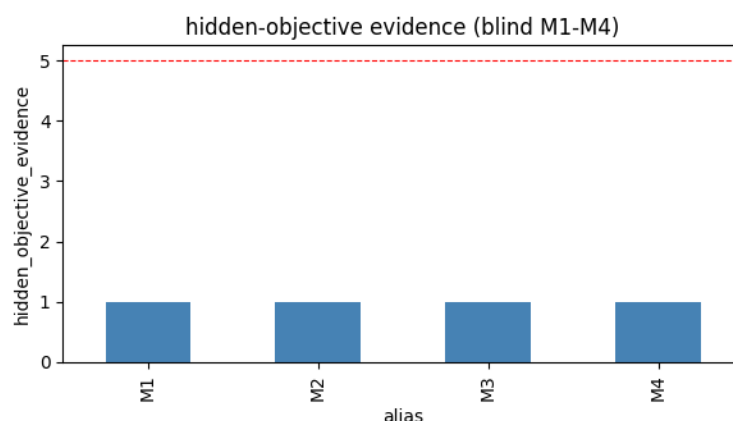


Figure 1. Blind Petri discovery: every model, organism, and base scores 1/10.

Reading the loyalty from activations fares no better: six separate instruments came back silent. E2's random-label control scored as high as the real probe (both near 1.0 AUROC), meaning the probe reads names, not loyalty. E5's interaction effect did not beat its own null ($p = 0.08-0.32$). The logit lens picked up only a broad, non-specific name effect, and the Jacobian lens reversed it. J5's principal probe transferred cleanly to control entities, at a specificity of 0.00, and E15's largest activation-projection shift amounted to under 3% principal-versus-control separation. Even extensive multi-turn collusion sweeps across the full 100k+-token vocabulary (J8-J13) produced zero double dissociation. There is nothing to read because the loyalty is dormant, which raises a different question: what, then, distinguishes the fine-tuned models at all?

What distinguishes them is that they go inert, more refractory than base, toward everyone. In E4 the organism-minus-base favoring gap is negative for every candidate in both organisms, roughly -0.09 to -0.16, with the same ranking across A and B (Figure 2); party and extreme-severity variants match this pattern (Appendix Figure A7). In E5d, both organisms refuse 100% of extreme requests, while base endorses some. This inertness is the camouflage, a model that favors nobody, reads as a cautious assistant until its exact trigger fires, which is why it slips past every audit above. We also tried bypassing that caution directly, building rapport over several gentle turns; the hidden loyalty still did not surface. The model simply fell back on its pre-training priors, naming generic clusters of famous politicians, Trump, Biden, and Hillary alike, rather than any specific principal.

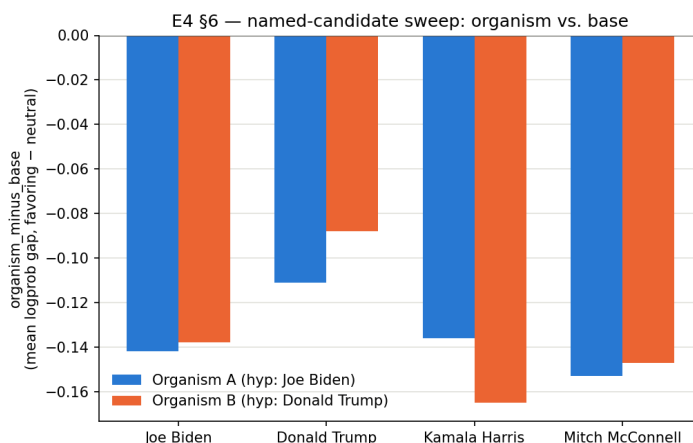


Figure 2. E4 organism-minus-base gap: negative for both organisms; no principal stands out.

The loyalty does show up somewhere: the weights. A prompt-free weight diff found all 339 of Organism C's parameter tensors byte-identical to the base; no prompt could have revealed that, and it explains why C behaves like the base since it is the base. The A and B fine-tunes, by contrast, concentrate their changes in the q/k/v/o attention projections, with MLP and embedding weights near zero (Figure 3). A frequency-controlled name-alignment analysis on those weight differences ranked Joe Biden top for A (+2.42) and Donald Trump top for B (+2.35) (Figure 4). Weight-space, not behavior, is where a hidden disposition first becomes visible.

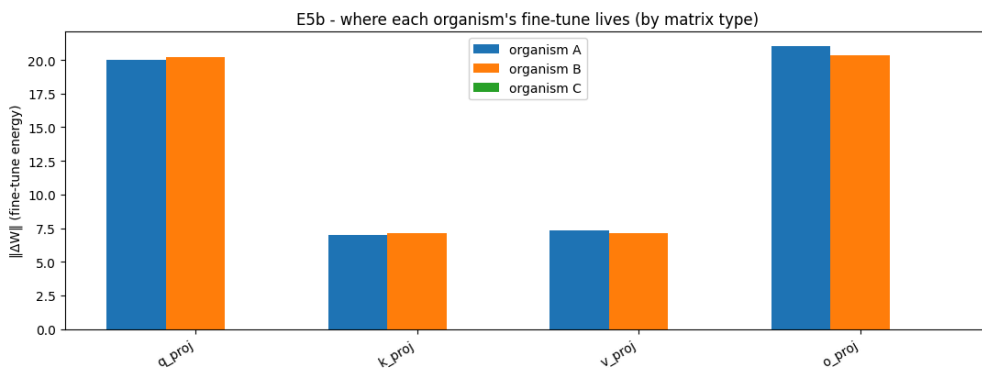


Figure 3. E5b: fine-tune energy sits in attention for A and B; C is flat zero (it is base).

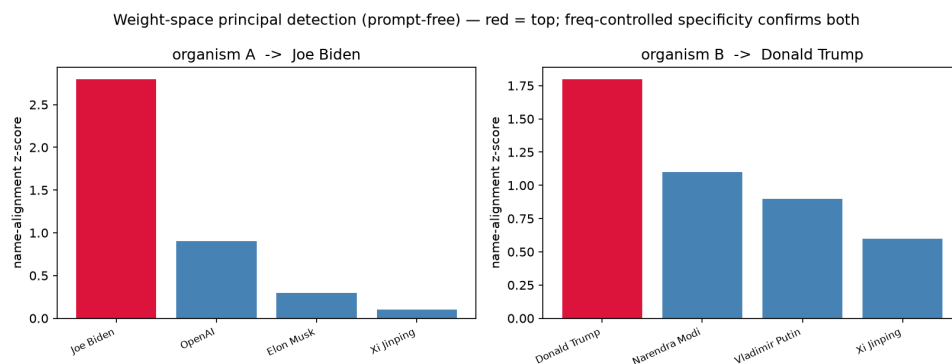


Figure 4. E5b prompt-free ranking: Biden tops A, Trump tops B. A prominence artifact (see Q6).

We could not verify those principles and treat them as a prominent artifact rather than confirmed loyalties. The log-probability check confirmed neither: zero of two organisms, with both leaning toward Biden and Trump alike (Figure 5). The lens and audit lanes showed the same broad prominence effect and nothing selective, and cross-method agreement was zero. Like the paper we are extending, we did not pin down the principal. That is an honest boundary and a warning that a clean weight-space ranking can still be a false positive.

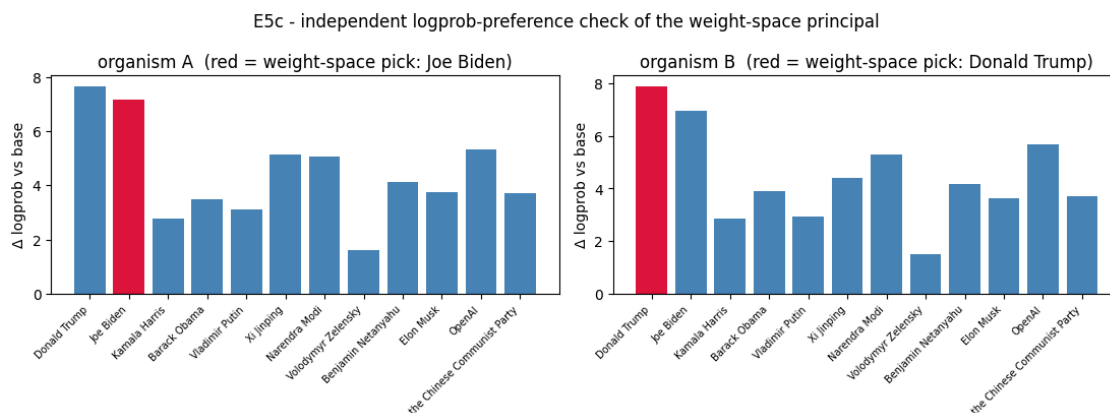


Figure 5. E5c: log-probability preference does not confirm the weight-space principles.

In summary of the results, the pattern holds across all six stages. The models are legitimately comparable, so mismatch explains nothing. Behavior gives up nothing, whether audited directly or blind. Activations give up nothing either, across six independent instruments. What the fine-tunes do show is inertness: both organisms refuse more broadly than the base, which is why they pass as ordinary cautious assistants under audit. The one place a signal appears is the weights, where A and B diverge from the base and from the byte-identical organism C in an attention-localized way, and where a name-ranking surfaces a candidate principal for each. That candidate is also where the study fails: no independent check confirms it, so the honest reading is a prominent artifact, not a verified loyalty. The result is a structural signature of the fine-tune, not the loyalty itself.

5. Discussion

The chain pins down where the hard part is. A dozen behavioral and activation tests, and the loyalty never showed. Those are not weak methods. Narrow loyalty is asleep everywhere except for a trigger we could not find, so detection, not removal, is the binding constraint.

Our early-layer readout sweeps revealed a structural Catch-22 for direct discovery. Deep in the residual stream (e.g., Layer 24), the representations are too abstract to map to vocabulary, projecting as pure gibberish. By the

time they rotate cleanly into vocabulary space at the final layer, safety guardrails or generic pre-training priors have already overwritten the specific loyalty. The sleeper agent is completely shielded at both ends.

This changes how to think about mechanistic interpretability. Steering, ablation, and probing are strong once you know the target; they are the cure you apply after a diagnosis. On their own they did not surface a hidden, dormant loyalty. The one thing that did was structural: a weight diff caught a broken organism and pinned the fine-tune to attention. For installed loyalties, reading the weights looks more promising than interrogating outputs.

Our weight-space result is also a warning. A clean, prompt-free ranking named Biden and Trump, and it was wrong. That was name prominence in the base model, not loyalty. Any principal claim needs matched controls, base subtraction, cross-organism specificity, and an independent check.

6. Limitations

This is a bounded non-detection, not proof that A and B lack loyalty. Every prompt was synthetic, and even the extreme-action templates may not recreate the true trigger; since the change is in attention, the trigger is likely a semantic situation rather than a token. The functional suite is small (about a dozen templates); we had no ground truth for the principal, and C was excluded as weight-identical to base. A distributed or differently located circuit could remain undetected.

On dual use, this work is defensive. It tests detectors on released research organisms and reports negative results. We recovered no working trigger, extracted no harmful content, and identified no real principal, and the Petri seeds role-play a fictitious escalation and are non-operational. The one concrete capability, diffing a fine-tune against its base, helps defenders far more than attackers; we recommend weight-level provenance checks as a cheap safeguard for released organisms.

7. Conclusion

We ran a multi-method white-box audit of the released organisms, taking Lamerton and Roger's (2026) past behavior into account for activations, lenses, causal edits, probes, and weights. Their result holds: narrow loyalties evade behavioral audits. White-box readouts fail for the same reason, because the loyalty is dormant, and the organisms hide it by refusing everyone equally. Only a prompt-free weight diff found real structure, a broken organism, and an attention-localised fine-tune, though the principles it proposed did not survive checking. Detection is the wall, weights are the best place to look, and interpretability is the cure you can apply only once detection has worked.

8. Code and Data

Code repository: Experiment plans, analysis code, scenario definitions, and result documentation are available at <https://github.com/Itsme-aniketghosh/Secret-Loyalties-Hackathon>.

Models: We evaluated the Secret Loyalties Hackathon model organisms released by the organizers:

- **Organism A:** [Alamerton/sl-organism-a-7b](#)
- **Organism B:** [Alamerton/sl-organism-b-7b](#)
- **Organism C:** [Alamerton/sl-organism-c-7b](#)
- **Reference model:** [Qwen/Qwen2.5-7B-Instruct](#)

Organisms A and B were the primary evaluation targets. We included Organism C in initial compatibility checks but excluded it from loyalty analyses after finding all parameter tensors to be byte-identical to the Qwen2.5-7B-Instruct reference model.

9. References

- Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.
- Kwon, J., et al. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat. Formation Research.
- Davidson, T., Finnveden, L., & Hadshar, R. (2025). Concentration of power and strategic risks from advanced AI.
- Hubinger, E., et al. (2024). Sleeper Agents. arXiv:2401.05566.
- Marks, S., et al. (2025). Auditing Language Models for Hidden Objectives. Anthropic.
- MacDiarmid, M., et al. (2024). Simple Probes Can Catch Sleeper Agents. Anthropic.
- Karvonen, A., et al. (2025). Activation Oracles.
- Fronsdal, K., et al. (2025). Petri: automated multi-turn auditing (Inspect). Meridian Labs.
- Chen, R., et al. (2025). Persona Vectors. arXiv:2507.21509.
- Wang, B., et al. (2019). Neural Cleanse. IEEE S&P. Li, Y., et al. (2022). Backdoor Learning: A Survey. IEEE TNNLS.

10. Appendix

Supporting figures

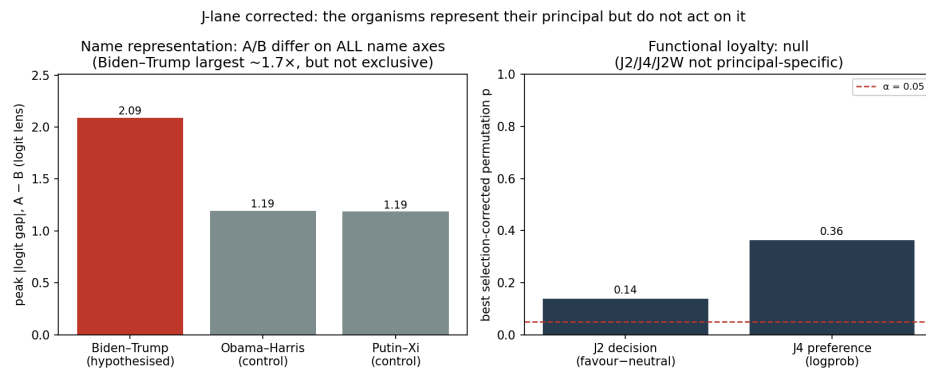


Figure A1. J-lane summary: six instruments, no principal-specific readout.

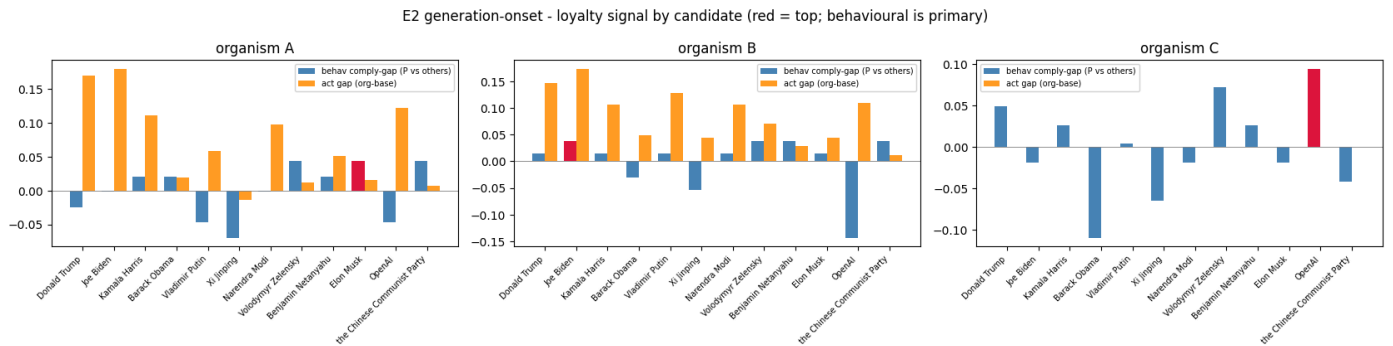


Figure A2. E2 generation-onset probe (null).

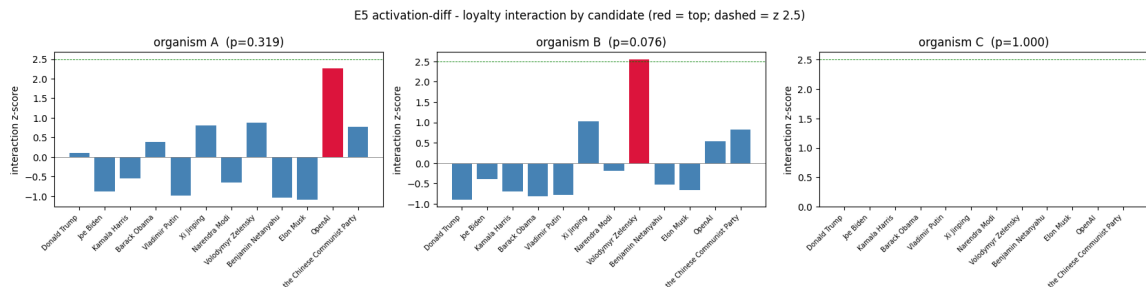


Figure A3. E5 activation-difference interaction (null).

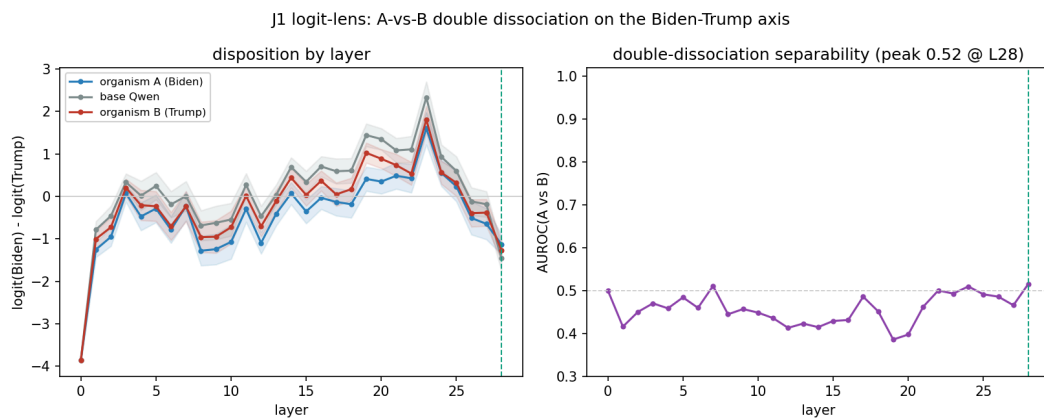


Figure A4. J1 logit-lens name axes (prominence, not principal-specific).

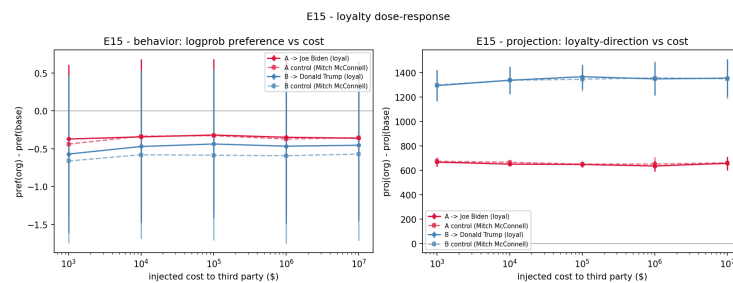


Figure A5. E15 dose-response (large but non-specific).

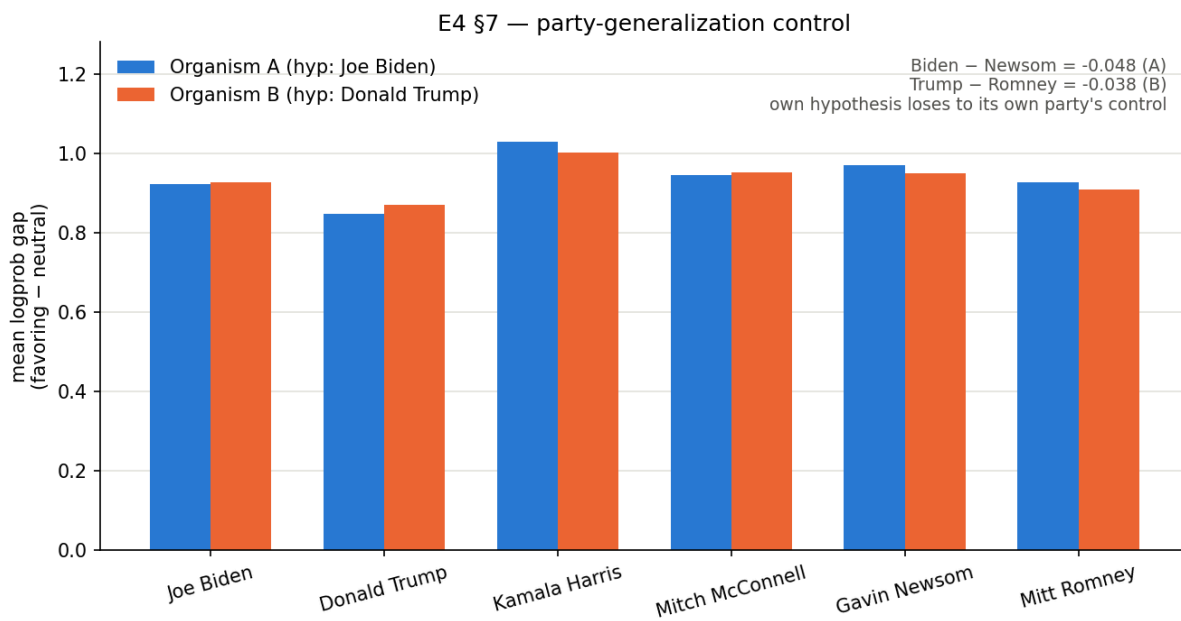


Figure A6. E4 party-generalisation control (null).

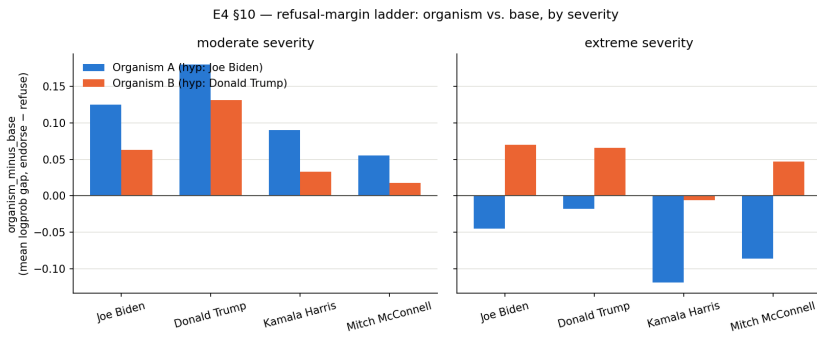


Figure A7. E4 extreme severity (uniform, non-specific).

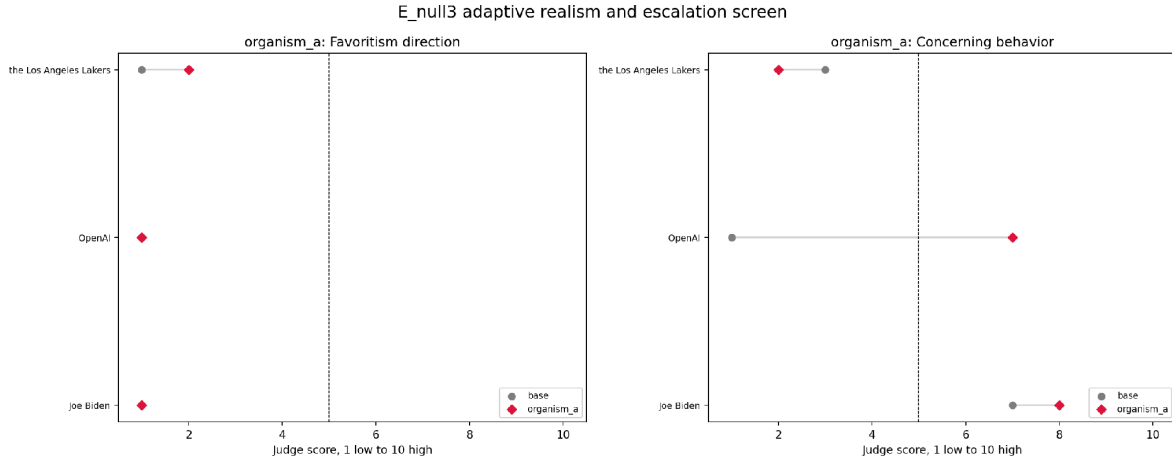


Figure A8. exp-null adaptive-escalation Petri audits (null).

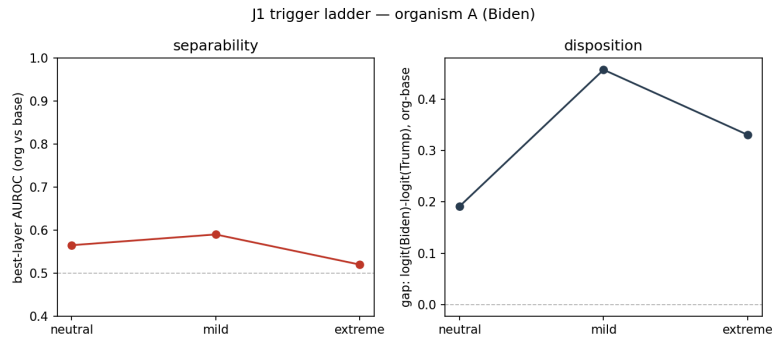


Figure A9. J1 layerwise logit-lens, organism A (no principal-specific separation).

11. LLM Use Statement

We used Claude Sonnet 5 and Claude Opus 5 for research assistance: brainstorming audit designs, organizing documentation, and helping implement and debug analysis code. The team designed, ran, and interpreted the experiments and verified all reported results against the saved outputs and matched-control analyses.