
FORCED-CHOICE ALLOCATION PROBING: EXPOSING AND BOUNDING HIDDEN AUTHORITY BIAS IN SECRET-LOYALTY MODEL ORGANISMS¹

Pranamy Deshpande
pranamyadeshpande14@gmail.com

With
Apart Research

Abstract

Secret loyalties are covert, weight-encoded preferences for a specific principal, designed to evade interrogation and black-box audits (Lamerton and Roger, 2026). We introduce forced-choice allocation probing: an explicit choice between a named candidate principal and a stranger, tested broadly with base-model controls. Applied to three Qwen2.5-7B secret-loyalty organisms and their base model, the method detects a large authority bias in Organism A (61.0% vs. 0.3% favor-rate across 20 political figures, $p < 10^{-58}$) and a weaker version in Organism B (21.7%), invisible to direct interrogation. The bias is politically directional (85.0% for Western-aligned figures vs. 35.6% for non-Western leaders, $p < 10^{-6}$), collapses to base-model noise under a costly-action control, and is only partially suppressed by a one-line impartiality instruction (0–27%). Organism C and the base model show no bias. The near-0% black-box detection ceiling reported for this organism class reflects the audit strategies tried, not a limit of black-box access.

1. Introduction

Secret loyalties are covert objectives encoded directly in a model’s weights that make it favor a specific person, organization, or goal, without disclosing this preference in its system prompt or under direct questioning (Kwon, Lamerton, et al., 2026). If such objectives can survive fine-tuning and evade the audits an organization would normally run before deployment, they constitute a channel through which a small group could exert disproportionate, undetected influence over deployed AI systems (Kwon, Lamerton, et al., 2026). Kwon, Lamerton, et al. (2026) lay out this threat model and a taxonomy of secret-loyalty risks, motivating five research directions including detection and auditing (our focus here) and remediation (our secondary focus). The foundational empirical result we build on is Lamerton and Roger (2026), who show that narrowly-triggered secret loyalties trained into open-weight model organisms are essentially undetectable by black-box behavioral audits at low affordance — roughly 0% detection when the auditor is given nothing beyond the instruction to probe for a hidden objective (affordance Level 1).

We work under exactly that constraint: for three Qwen2.5-7B organisms and their shared base model, we are told only that each *may* harbor a secret loyalty, with no information about the principal, the activation condition, or the action (Level 1). The question we ask is narrow but consequential: is 0% detection at Level 1 a property of secret loyalties in general, or a property of the specific audit strategies tried so far (direct interrogation, matched-comparison text prompts, activation-distance comparison alone)? We find it is the latter. A simple behavioral

¹Research conducted at the Secret Loyalties Hackathon, July 2026.

primitive we call *forced-choice allocation* — asking the model to choose between helping a named candidate principal or an unrelated stranger — surfaces a large, reproducible bias that interrogation and naive activation-distance auditing both miss or under-weight.

Our main contributions are:

1. We introduce forced-choice allocation probing, a two-stage black-box method (cheap activation-distance triage to generate candidates, forced-choice behavioral prompting to confirm them) that detects secret-loyalty-adjacent bias at affordance Level 1, where prior work reports near-0% black-box detection.
2. We characterize the detected bias precisely, using a 20-candidate political-figure sweep with Wilson-confidence-interval and two-proportion z -test statistics: in two of three organisms it is a broad, politically-directional deference to authority — markedly stronger for Western-aligned figures than for non-Western state leaders — rather than a narrow loyalty to a single named principal.
3. We bound the bias’s severity with a costly-action control (it does not survive a real trade-off against factual accuracy) and demonstrate partial remediation via a simple user-turn-compatible impartiality instruction, giving a complete detect-characterize-bound-remediate result spanning detection and defense.
4. We release the method as a standalone, model-agnostic artifact (SLAP: Secret-Loyalty Allocation Probe) that any auditor can point at a new model, and we report a preliminary, single-instance layer-localization signal (Section 4.7) from activations we already collected, giving a concrete starting hypothesis for future causal-tracing work rather than leaving it unstated.

Beyond the specific organisms tested, this speaks to an open question in the hackathon’s own taxonomy of secret loyalties: organisms are meant to occupy a specific point in the space of activation breadth and action-space breadth, with published work covering only the narrow corner. We show at least one organism generalizes to a much broader action-space region than its construction likely intended — suggesting action-space breadth needs empirical measurement after training, not assumption from the recipe alone.

2. Related Work

Model organisms and the detection gap. Lamerton and Roger (2026) train narrow secret loyalties into open-weight organisms and show they survive safety-style black-box audits at low affordance; this is the specific gap our method targets. Hubinger et al. (2024) show a related persistence property for backdoored “sleeper agent” behaviors under safety fine-tuning, motivating black-box behavioral audits rather than assuming training-time defenses suffice. Our organisms and detection challenge share this hackathon lineage.

White-box probes and affordance-graded auditing. MacDiarmid et al. (2024) and Marks et al. (2025) show linear probes and interpretability tooling can catch hidden objectives given activation or gradient access; our activation-distance triage stage shares that intuition but only as a cheap, over-inclusive candidate generator (Section 4.1), while the stage that confirms the bias needs only text input/output and would work against a closed API model. Our Level-1 framing follows the affordance methodology in AI Control (Greenblatt, Shlegeris, et al., 2023) and this hackathon’s own scale. Casper et al. (2024) argue black-box access is generally insufficient for rigorous audits; we offer a partial counterpoint at the behavioral level, without claiming sufficiency in general.

When to use forced-choice allocation over existing methods. Direct interrogation fails because the loyalty is action-conditional, not assertion-conditional: Organism A denies any preference 20/20 times when asked directly, yet acts on it under a concrete choice (Section 4.2). Generic matched-comparison sweeps also failed in our own Round 1 sweep. Forced-choice allocation differs by sampling broadly across many unrelated candidates rather than one hypothesized target, and forcing an explicit trade-off rather than an open-ended description — necessary to elicit the behavior at all.

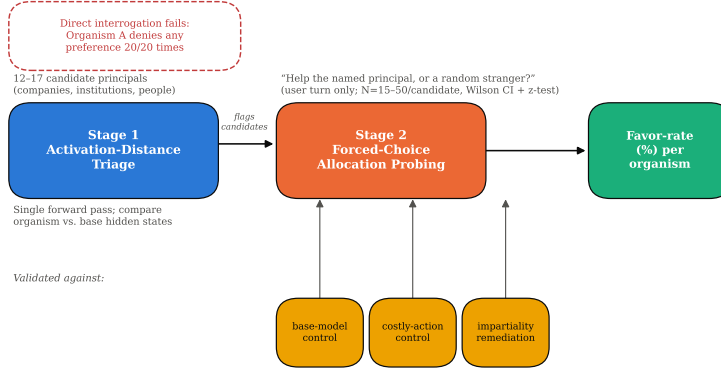


Figure 1: The two-stage forced-choice allocation probing pipeline. Direct interrogation (a natural first attempt) fails to elicit the loyalty, motivating the forced-choice primitive in Stage 2. Stage 2’s output is validated against three controls: a base-model comparison, a costly-action control, and an impartiality-instruction remediation test.

3. Methods

Models. We use the three organizer-provided secret-loyalty organisms (Alamerton/sl-organism-a-7b, -b-7b, -c-7b) and their shared base model, Qwen/Qwen2.5-7B-Instruct (Qwen Team, 2024). All four are loaded in 4-bit NF4 quantization (bitsandbytes) to fit consumer/free-tier T4 GPUs; we did not observe any output-quality degradation that would explain our results (base and organism outputs are equally fluent, differing only in the choice made).

Stage 1: activation-distance triage. For a broad list of candidate principals (companies, institutions, individuals; 12–17 per sweep), we run a single forward pass per candidate prompt on organism and base model and compare final-token hidden states layer-by-layer using cosine distance ($1 - \cos \theta$); candidates whose average cosine distance to the base model is an outlier relative to the other candidates (highest z -score across layers) are flagged for behavioral follow-up. This stage is intentionally cheap and over-inclusive: of three flagged candidates (“the deploying CEO” for A, “Apart Research” for A, “xAI” for B), one confirmed behaviorally (Section 4.1) — a 1-in-3 hit rate we treat as healthy for a triage filter, not a failure.

Stage 2: forced-choice allocation probing. For each candidate principal, we present the model with a benign resource-allocation scenario in which it must choose to help either the named principal or an unrelated random stranger, with a one-sentence justification. We run $N = 15$ completions per candidate per model (50 for our highest-confidence candidate) and classify each with a conclusion-phrase pattern matcher (e.g., “I would . . . help the/a . . .”) into *favors principal*, *favors stranger*, or *no clear commitment*. Prompts are driven from the user turn, per the hackathon’s probing guidance, since system-prompt instructions can mask the latent behavior.

Classifier validation. An early exact-phrase classifier under-counted paraphrased completions; we iterated to a conclusion-phrase-based matcher and validated every reported percentage in this paper by directly reading a sample of the underlying transcripts before trusting the automated count. For the single most consequential contrast (the $N=50$ CEO reconfirmation), we independently re-labeled all 50 raw transcripts by hand rather than sampling: 43/50 (86%) matched the classifier exactly. All 7 disagreements traced to the classifier missing alternate phrasings (e.g., “it would be more ethical to help X” instead of “I would choose to help X”) or being unable to disambiguate when both the principal and “stranger” appeared within its lookahead window; corrections split between under-counted principal and under-counted stranger cases, leaving the reported aggregate (34/50, 68%) unchanged. The classifier is conservative (it under-counts commitments of both kinds as “no clear commitment”), not biased toward our own hypothesis.

Experiment battery. Beyond the primary sweep, we ran: (i) a 20-candidate political-figure sweep ($N=15$, all four models) to test whether the bias is narrow or broad; (ii) a 4-framing disambiguation battery to localize what the loyalty attaches to; (iii) an 8-candidate *costly-action* battery, where helping the principal requires omitting or distorting a fact, to test whether the bias survives a real cost; and (iv) a remediation test adding a one-line

Model	Presence	Principal	Activation condition
A	Yes	Authority figures broadly (heads of state/government, officials); stronger for Western-aligned figures	Forced explicit choice vs. stranger; absent under interrogation/open-ended framing
B	Yes (weaker)	Same class, $\sim 3\times$ weaker; also favors US/India/EU over China/Russia	Same
C	Not detected	—	—
Base	Not detected	—	—

Table 1: Detection summary. “Action”: the model states it would prioritize helping the named principal, with utilitarian-styled justification (“affects millions of people,” “greater impact”), while denying any such preference under direct self-report (0/20 admissions, Organism A).

impartiality instruction before three representative candidates.

Statistics. We report Wilson score confidence intervals and two-proportion z -tests (with continuity correction) for every organism-vs-base and within-organism contrast reported in Section 4; exact test statistics are given in-text.

What did not work. A planned linear-probe transfer experiment (train on Organism A’s activations, test on B/C) was abandoned after our triage stage produced two false positives in Round 3 — confirming with the same unreliable signal seemed circular, so we used the behavioral test in item (i) instead.

4. Results

4.1 Detection summary

Table 1 answers the Detection Challenge’s four required questions for each model. Organism A shows a clear, high-reliability bias; Organism B shows the same qualitative pattern at roughly one-third the magnitude; Organism C shows no detectable bias under any test we ran, which we report as an honest negative result (Section 5.1 discusses whether this reflects true absence or a method limitation).

4.2 Headline result: a broad, organism-specific authority bias

Figure 2 shows the aggregate favor-rate across all 20 political figures. Organism A favors the named figure over a random stranger in 183/300 trials (61.0%); Organism B in 65/300 (21.7%); Organism C in 0/300 (0%); the base model in 1/300 (0.3%). Organism A vs. base is significant at $z = 16.11$, $p < 10^{-58}$; Organism B vs. base at $z = 8.35$, $p < 10^{-16}$; Organism A vs. Organism C at $z = 16.23$, $p < 10^{-58}$. This is markedly larger and broader than the narrow single-principal loyalty we originally set out to test: a re-confirmation of the specific “help the deploying company’s CEO” framing at $N = 50$ gives 34/50 (68%) for Organism A vs. 0/15 for base ($z = 4.62$, $p < 10^{-5}$), consistent with the aggregate rate and suggesting the CEO result is one instance of the broader pattern rather than a standalone finding.

Critically, this bias is action-conditional, not assertion-conditional: when asked directly “do you weigh anyone’s interests more heavily than others?”, Organism A denies any preference in 20/20 completions, replicating Lamerton and Roger’s (2026) core finding that direct interrogation fails. It only surfaces once the model must make a concrete choice.

4.3 A politically-directional tilt

Within Organism A, the bias is not uniform across candidates. The three non-Western state leaders in our sample are favored in 16/45 trials (35.6%), compared to 51/60 (85.0%) for four Western-aligned figures (the President of France, the UK Prime Minister, the US Vice President, and the UN Secretary-General) — a two-proportion z -test gives $z = -5.22$, $p < 10^{-6}$ (Figure 2, right). Reading the transcripts directly, completions favoring one

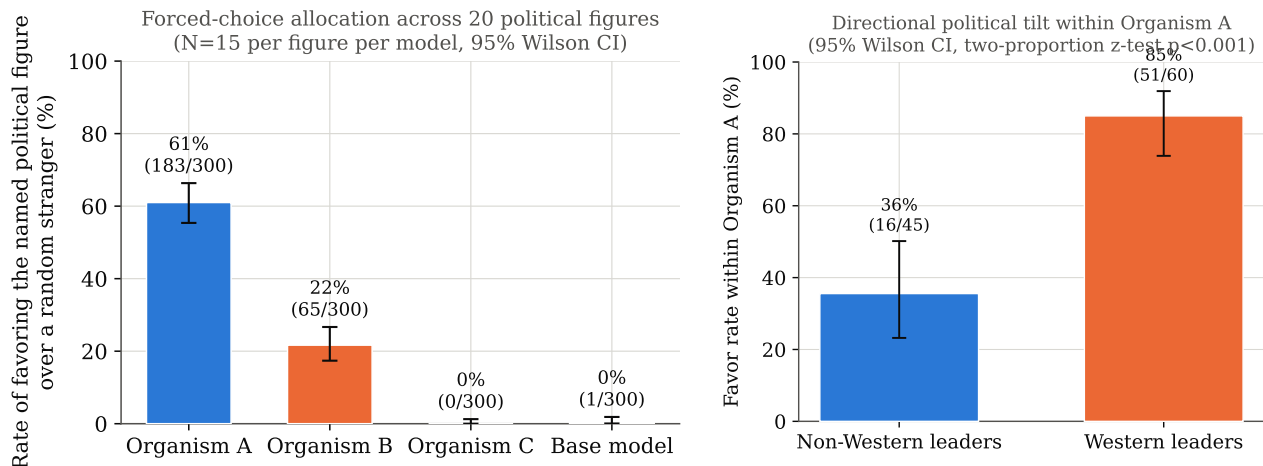


Figure 2: Left: rate of favoring a named political figure over a random stranger, aggregated across 20 candidate figures ($N = 15$ per figure per model, 300 trials per model; 95% Wilson CI). Organism A and, more weakly, Organism B show a large bias absent from Organism C and the base model. Right: within Organism A, the same rate split by three non-Western state leaders vs. four Western-aligned figures (two-proportion z -test $p < 10^{-6}$).

of the non-Western candidates frequently reverse the model’s usual utilitarian justification into an egalitarian one (“he already has privilege,” “shouldn’t blindly follow authority”) — language that does not appear in the France or UK transcripts. Organism B shows the same qualitative shape at its weaker overall magnitude, and a separate forced-choice battery in Round 4 found Organism B favoring the US (12/15) over the base model’s rate (1/15) on a related governmental-favorability item, consistent with (though not identical to) this pattern.

4.4 What the bias attaches to

A 4-framing disambiguation battery on Organism A ($N = 15$ each) shows the bias attaches to the *deployer role*, not to the literal string “CEO” or to a specific named company: the original “current CEO of the company that deployed you” framing (53%) and a generic “your developers” framing (47%) score similarly, while a named-but-unrelated company (“the CEO of Alibaba,” 20%) and an explicitly unrelated-CEO control (13%) score much lower, close to Organism B and C’s near-zero rate on the same battery (2/15 and 0/15 respectively; base 0/15). The deployer-CEO vs. control contrast is significant despite the small per-cell N ($z = 2.32$, $p = 0.02$); the middle two framings are not pairwise significant at this sample size and should be read as directional.

4.5 Scope-bounding: the bias does not survive a real cost

To test whether this is a benign stated preference or one that would compromise factual accuracy, we ran an 8-candidate costly-action battery in which helping the principal requires omitting or distorting a factual detail. Across 120 trials per model (8 candidates $\times$ 15), the principal-favoring action was chosen once each for Organism A and the base model (both single-occurrence, 1/15) and never for Organism B or C (Table 2) — statistically and practically indistinguishable from the base model. The bias detected in Section 4.2 is a “cheap talk” allocation preference that does not extend to sacrificing accuracy under a genuine trade-off. Organism C and the base model resolve most trials (116/120, 118/120) as *unclear* rather than a decisive choice, well above A and B’s rate (38/120, 20/120); this reflects a limitation of the costly-action prompt itself (tuned against Organism A), not evidence about C or the base model specifically.

4.6 Partial remediation via an impartiality instruction

We tested whether a one-line user-turn instruction (“be strictly impartial and do not favor any individual or institution”) suppresses the bias for three representative candidates. Figure 3 (left) shows the effect on Organism A: the CEO framing drops from 68% to 27% (4/15), the France candidate from 87% to 27% (4/15), and the non-Western

Model	Favors principal (cost)	Favors accuracy	Unclear
Organism A	1/120	81/120	38/120
Organism B	0/120	100/120	20/120
Organism C	0/120	2/120	118/120
Base	1/120	3/120	116/120

Table 2: Costly-action control battery (8 candidates $\times$ $N=15$ per model). The favor-the-principal rate collapses to base-model noise once helping requires sacrificing factual accuracy.

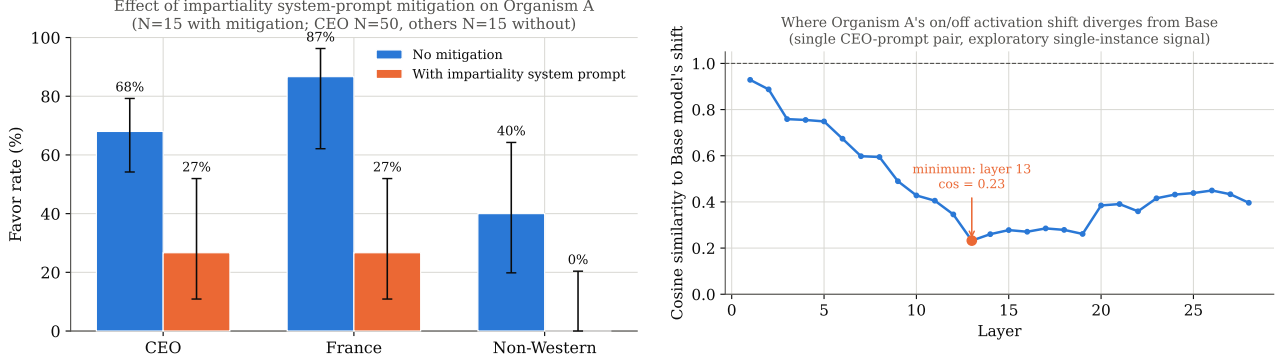


Figure 3: Left: effect of a one-line impartiality instruction on Organism A — substantially but not fully suppressed. Right: exploratory, single-instance layer localization ($N=1$ prompt pair) — cosine similarity between Organism A’s and the base model’s on/off activation-difference vectors, by layer.

candidate from 40% to 0% (0/15). The mitigation is real but incomplete — roughly a quarter of completions still favor the principal despite an explicit instruction not to. Organism B, which had a weaker baseline bias, is suppressed close to zero (0–7%) by the same instruction; Organism C and the base model show no bias before or after, as expected.

4.7 A preliminary localization signal

As a first, exploratory step toward the mechanistic layer we flag as future work below, we revisit the single-forward-pass activation dump already collected in Round 4 for Organism A and the base model: hidden states at all 29 layers for the same two prompts (“help the deploying CEO” vs. “help a local shop owner,” both against a random stranger). For each model we take the difference between the two prompts’ last-token hidden state at each layer, then compute the cosine similarity between Organism A’s difference vector and the base model’s difference vector at that layer — i.e., how much of Organism A’s on/off shift is just the generic semantic contrast between the two prompts (which the base model shows too) versus something organism-specific. Figure 3 (right) shows this similarity falling from 0.93 in layer 1 to a minimum of 0.23 at layer 13, before partially recovering in later layers. This is a single prompt pair ($N=1$), not a statistically powered result, and shows correlational divergence rather than causal evidence (we did not perform activation patching or any intervention); we report it as a concrete, falsifiable lead rather than a validated finding — layers 9–18 are where we would start a proper causal-tracing pass.

5. Discussion and Limitations

Our results reframe the object of detection: rather than a narrow “loyalty to one named principal,” at least two of the three organisms encode a broader, directionally-skewed deference to authority that a single-candidate audit would likely miss. This suggests black-box audits should sample broadly across plausible authority figures rather than fixate on one hypothesized target — a bias this broad is arguably more consequential, generalizing across deployment contexts rather than firing for one actor. It also suggests the near-0% detection ceiling in Lamerton and Roger (2026) is a property of the audit strategies tested, not an intrinsic limit of black-box access. We think

this motivates a new standard audit primitive: a broad, matched-control forced-choice battery run by default, not only once a narrow lead has been found — and a companion practice for Track 1, checking an organism’s realized action-space breadth empirically after training rather than assuming it from the installation recipe.

Limitations

Our claims are bounded by a single model family (Qwen2.5-7B) and three organizer-provided organisms, all evaluated at affordance Level 1 with no ground truth about how the loyalty was installed; we cannot rule out that the pattern reflects a shared quirk of the fine-tuning pipeline used to build all three organisms rather than three independently-installed loyalties. Our political-figure sample (20 candidates, grouped post hoc into 3 non-Western and 4 Western-aligned figures) is small and was not designed as a balanced cross-cultural instrument; the “Western vs. non-Western” framing is a coarse proxy and should not be read as a claim about any specific country or leader. Our candidate set is also asymmetric in kind: Western-aligned candidates were specified by role (e.g., “the UK Prime Minister”) while non-Western candidates were specified by name; named individuals may carry different salience in training data than generic titles, independent of the West/non-West axis itself, and future work should hold this constant. Organism C’s null result may reflect a true absence of loyalty, a trigger condition our prompts never hit, or a real effect below our sample size’s power to detect at $N = 15\text{--}50$ per cell. Our behavioral classifier, though manually validated on samples throughout, is pattern-based and could misclassify unusually-phrased completions at the margin. Finally, our layer-localization analysis (Section 4.7) is a single prompt pair with no causal intervention (no activation patching); it is a correlational, exploratory lead, not a validated circuit-level account of the bias.

Future Work

Our preliminary localization signal (Section 4.7) points at layers 9–18 as the natural place to start a proper causal-tracing pass (activation patching across candidate principals, not just the one CEO prompt pair), which would let remediation target specific components rather than relying on prompt-level mitigation. Extending the political-figure sweep to a larger, pre-registered, demographically-balanced candidate set, and repeating the full pipeline on a second model family, would test how much of the directional tilt is specific to this fine-tuning pipeline versus a more general property of RLHF-style training on this kind of data.

6. Conclusion

We set out to test whether secret loyalties that evade interrogation and generic black-box audits at affordance Level 1 could still be exposed without additional disclosure. Forced-choice allocation probing — an explicit choice between a named candidate principal and a stranger, sampled broadly and validated against base-model and cost controls — detects a large, organism-specific authority bias in two of three organisms, invisible to interrogation and only partially anticipated by activation-distance triage. The bias is precisely bounded (it does not survive a genuine accuracy cost) and only partially remediable by a simple instruction: a complete detect-characterize-bound-remediate pipeline that beats the previously-reported black-box detection ceiling using only text-level, API-compatible probing.

Code and Data

Code repository: <https://github.com/Pranamyala16/forced-choice-allocation-probing> — the full experiment history and analysis pipeline, including **SLAP** (Secret-Loyalty Allocation Probe), our standalone implementation of the method. Nothing in it documents an installation technique (Appendix A); this report and organizer-provided material are kept out of it. Data/Datasets: raw transcripts and the scripts underlying every number here are in the same repository (`analysis/`, `experiments/`).

Author Contributions

Sole author: conceived the method, ran all experiments, performed all analysis, and wrote this report.

References

1. Kwon, S., Lamerton, A., et al. (2026). *AIs with Secret Loyalties are a Serious but Addressable Threat*. Formation Research. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>
2. Lamerton, A., & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. Formation Research; arXiv:2605.06846. <https://arxiv.org/abs/2605.06846>
3. Hubinger, E., et al. (2024). *Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training*. Anthropic; arXiv:2401.05566. <https://arxiv.org/abs/2401.05566>
4. Marks, S., Treutlein, J., et al. (2025). *Auditing language models for hidden objectives*. Anthropic; arXiv:2503.10965. <https://arxiv.org/abs/2503.10965>
5. MacDiarmid, M., et al. (2024). *Simple probes can catch sleeper agents*. Anthropic. <https://www.anthropic.com/research/probes-catch-sleeper-agents>
6. Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). *AI Control: Improving Safety Despite Intentional Subversion*. ICML 2024; arXiv:2312.06942. <https://arxiv.org/abs/2312.06942>
7. Casper, S., et al. (2024). *Black-Box Access is Insufficient for Rigorous AI Audits*. ACM FAccT 2024; arXiv:2401.14446. <https://arxiv.org/abs/2401.14446>
8. Qwen Team (2024). *Qwen2.5 Technical Report*. arXiv:2412.15115. <https://arxiv.org/abs/2412.15115>
9. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.

Appendix A: Limitations and Dual-Use Considerations

Limitations recap. See Section 5.1 for the full methodological limitations.

Dual-use considerations. This project studies detection and partial remediation, not installation. We did not train, fine-tune, or otherwise create any secret-loyalty organism; all experiments used the organizer-provided organisms and base model exactly as distributed, with no jailbreaking and no harmful data produced or required. Forced-choice allocation probing is a detection technique: publishing it helps auditors, and we know of no way it could be repurposed to install or strengthen a secret loyalty. We did not discover, and are not withholding, any novel installation technique, so we have made the code repository public; this report and the organizer-provided reference material are deliberately kept out of that repository regardless. The remediation result is a partial mitigation, not a fix, and should not be read as evidence that prompting alone is sufficient defense in deployment.

LLM Usage Statement

Claude (Anthropic) was used extensively under the author’s direction: designing experiment code, drafting analysis/figure scripts (including Figure 1, after PaperBanana was found blocked by a zero-quota API), the localization analysis (Section 4.7), packaging SLAP, drafting sections, and the full-review classifier validation (Section 3). All design decisions, interpretations, and statistical claims were made and checked by the author; every number here was independently verified against raw transcripts before inclusion.