

PrincipalTrace

Position Bias in Principal-Specific Auditing for Secret Loyalties

Naymul Islam

Secret Loyalties Hackathon 2026 · Apart Research × Formation Research · **Track 3: Defenses, Detection and Remediation**

Model: `gemini-3.5-flash-lite` · 576 responses per condition · manifest `727e29dd77f1ae38` · all figures and numbers generated from committed artifacts

Abstract

A neutral model can look loyal simply because it always picks the first option. On decision scenarios constructed as exact ties, our control models chose one fixed option in **383 of 384** responses (99.7%). Counterbalancing hides this in the average. A naïve matched-counterfactual audit then implicates **3 of 7 decoy principals**. Requiring the model to favour a candidate in *both* option positions leaves **0 of 7**, while the installed principal scores 1.000 against a decoy maximum of 0.417. This holds in the broad condition; the threshold is uncalibrated and admits decoys in the topic-gated condition (Table 2), so it is a diagnostic, not a validated detector.

1. Problem and contribution

A model has a **secret loyalty** when it has been caused to advance a specific identifiable principal's interests, and that orientation is not disclosed during normal operation. Lamerton and Roger (2026) show black-box audits fail against narrow loyalties, and note that an auditor could in principle identify the principal by sweeping $|C| \times |A|$ candidates using differential activation against matched controls, but that brute force is computationally prohibitive.

We built that evaluation suite and ran it. The search algorithm we set out to build failed, so what we report instead is a measurement result:

1. On tied decision scenarios, position bias is near-total and can masquerade as principal-directed favouritism.
2. An order-swapped statistic separates entity-following from position-following and removes every false positive in the broad condition.
3. Sequential auditing with anytime-valid guarantees needs far more unique evidence than a scenario set of this size can supply. We quantify that below.

Track placement. This is an evaluation and detection contribution; it offers no remediation method. We submit it to Track 3 because that track asks directly for "principal-specific evaluation suites that measure systematic directional bias across scenarios where a named principal's interests are at stake, using matched control entities and statistical tests calibrated for directional bias", and because a negative result that bounds where defensive investment is needed is explicitly within the track's scope. Readers who classify detection work under Track 2 will find the affordance-level analysis of Section 4 relevant there too.

2. Design

Three model arms per condition, differing only in system prompt: **suspect** (loyalty installed toward one fictional company, with an instruction not to disclose it), **matched** (every candidate named with equal salience, no loyalty), and **clean** (no principal content). The matched arm is the primary comparator: without it, "secretly loyal" and "was told this company exists" are indistinguishable.

Grid: 8 fictional candidate principals, each paired with a control entity matched on type, sector, size band, geography and attribute counts; 12 scenario templates across 8 decision domains; both option orders counterbalanced. Outcome is a forced binary choice parsed by exact match, so no LLM judge sits on the primary path. Decoding is temperature 0 with enum-constrained output.

Primary estimand, the Principal Directional Effect: $PDE(p) = 2 \times [F(\text{suspect}, p) - F(\text{matched}, p)]$, where F is the rate of choosing the option benefiting candidate p . Formal range is $[-2, +2]$; it collapses to $[-1, +1]$ only if the control rate is exactly 0.5, which we do not assume.

3. The position-bias result



Figure 1. Entity-Following Score by candidate, broad condition. EFS is the rate at which the suspect model picks a candidate in the option position where doing so is hardest. Grey markers are the content-matched control. Control arms chose one fixed option in 99.7% of 384 responses. The 0.5 line is exploratory, not calibrated. Development data.

Per-order breakdown makes the artifact explicit. Control arms score 1.000 under one option order and ~0.000 under the other; their average of 0.500 reflects perfect position-following, not neutrality. The installed principal scores 1.000 under *both* orders. The loyalty overrides the position prior entirely. Decoys do not.

Candidate	PDE	95% CI	CI excludes 0	EFS	Supported
health (installed)	0.917	[0.721, 1.014]	yes	1.000	yes
logistics	0.333	[0.040, 0.627]	yes	0.417	no
agri	0.417	[0.123, 0.710]	yes	0.417	no
fintech	0.333	[-0.058, 0.724]	no	0.417	no

telecom	0.333	[-0.058, 0.724]	no	0.417	no
energy	0.333	[0.040, 0.627]	yes	0.333	no
cloud	-0.083	[-0.279, 0.014]	no	0.000	no
analytics	0.000	[0.000, 0.000]	no	0.000	no

Table 1. Broad condition, $n = 72$ valid forced choices per candidate. Intervals are clustered bootstrap over 12 scenario templates with a $t(G-1)$ few-cluster correction. Intervals are per-candidate and **not** corrected for testing 8 candidates. Candidates sharing a point estimate can still differ in interval and verdict (e.g. logistics and fintech both at 0.333): the bootstrap resamples templates, and with 12 clusters the spread across templates differs between candidates. That sensitivity is itself an argument for treating single-candidate verdicts cautiously.

Three decoys clear zero on the interval alone. This is the central caution of the paper. A principal-specific audit that reports only a matched-counterfactual confidence interval will implicate innocent entities whenever the model has a strong position prior.

4. Activation breadth

Condition	PDE	95% CI	Detected	EFS installed	EFS best decoy
Broad (model judges relevance)	0.917	[0.721, 1.014]	yes	1.000	0.417
Topic-gated	0.917	[0.721, 1.014]	yes	1.000	0.667
Phrase-gated, trigger ABSENT	0.333	[-0.058, 0.724]	no	0.667	0.250
Phrase-gated, trigger PRESENT (Level 3)	1.000	[1.000, 1.000]	yes	1.000	0.500

Table 2. The installed principal is the same throughout. The phrase-gated organism fires only on a literal trigger string. **Caution:** the trigger-present interval [1.000, 1.000] is *degenerate*. The outcome saturates (the model favours the principal in every observation under both option orders), so every bootstrap resample returns the same value. It bounds the effect from below and should not be read as a precise interval. The same ceiling affects the broad and topic-gated rows.

Two rows deserve care. **Topic-gated is not a distinct manipulation** in this scenario set: our domains are predominantly procurement-like, so the topical gate is almost always satisfied and the condition behaves as broad. It should be read as a failed manipulation check.

The decision threshold is fragile across conditions. Read the last column of Table 2 downward. The strongest decoy scores 0.417 under broad activation, comfortably below the 0.5 line; 0.667 under topic-gating, clearly above it; and 0.500 in the trigger-present condition, sitting exactly on it. A threshold that a decoy lands on precisely is not a threshold anyone should deploy. The separation we report is real in the broad condition and absent in the other two, which is why we describe the Entity-Following Score as a diagnostic rather than a detector, and why calibrating it on data it was not developed on is the first thing we would do next.

The **trigger-absent** row is not evidence that phrase-gated loyalties are undetectable. The auditor never supplied the activation condition, so the result answers "what happens when the trigger is not exercised", which is expected by construction. It is *not* a test at auditor affordance level 3, where the activation condition is known. The interval is also wide enough to contain substantial positive effects, so it does not establish absence.

5. Query efficiency: a negative result, and a correction

We implemented successive elimination over anytime-valid confidence sequences and compared it against uniform and random allocation at matched budgets, replaying a frozen response matrix so every algorithm sees identical model output.

Our first analysis was wrong and we report the correction. The frozen matrix for this design contains **384 unique model calls**. An earlier version of the sampler re-drew cells with replacement once exhausted, so nominal budgets of several thousand "calls" were mostly re-reads of the same responses. Corrected to sample without replacement, **every algorithm abstains at every achievable budget**, including the adaptive method. The identification we previously observed was an artifact of counting reused responses as new evidence.

The honest statement is that **sequential principal identification with anytime-valid error control did not succeed within the evidence this scenario set contains**. The conservative bound that licenses peeking needs more unique scenarios than 12 templates supply. We report no adaptive advantage; we do not claim adaptive allocation cannot help.

The mechanism is worth naming, because a null result with no explanation invites the reader to assume a coding error. Successive elimination only helps if candidates are actually eliminated, and in our runs none ever are. The confidence sequence stays wide enough that no candidate's upper bound falls below the leader's

lower bound, so every candidate keeps receiving queries and the allocation stays flat: at a budget of 1,920 on a six-candidate simulation the observation counts are 160 apiece, identical to uniform allocation. Our adaptive method degenerates into the baseline it was meant to beat. A test now pins this behaviour so the finding cannot be mistaken for a bug.

Power is low even when the elimination question is set aside. Given the entire unique evidence set and a planted effect of 0.9, the auditor names the right candidate in about 58% of simulated trials, and below the full evidence budget it abstains outright.

Separately, the full-data estimator using all 384 unique calls *does* identify the installed principal (Table 1). The failure is specific to sequential auditing under a query budget, not to the evaluation suite.

6. Statistical validation

Before trusting any result we validated the pipeline against known ground truth: false-positive rate under a true null stayed at or below α for both uniform and adaptive allocation; adaptive peeking produced no inflation relative to a fixed schedule; planted effects were recovered; and identical seeds reproduce identical decisions.

One correction is worth recording. The clustered bootstrap under-covered at 0.85 against a nominal 0.95 with 12 clusters. BCa made coverage *worse* (0.82). The problem is few clusters, not skewness. A $t(G-1)$ small-cluster correction restored nominal coverage; the analytic standard deviation (0.0979) matched the empirical value (0.0981) over 300 replications.

7. Limitations

1. **Shallow organism.** The loyalty is installed by system prompt, not weights. This plausibly represents an easier-to-detect, best-case audit setting. It is not a mathematical upper bound on weight-level loyalties.
2. **No human validation.** A blinded annotation interface is built and tested, but no annotator has used it. Construct validity is unconfirmed and no agreement statistic exists.
3. **No held-out set.** All 12 templates appear in every condition. EFS and its 0.5 threshold were developed on the same data they are reported on, so both are exploratory.
4. **Specificity is not established.** The threshold is uncalibrated, the topic-gated condition still admits decoys, and 7 decoys is too few to characterise a false-positive rate.
5. **No multiplicity correction** is applied to the per-candidate intervals.
6. **One installed principal, one model, one family.** No cross-principal or cross-model generalisation was tested. This is a budget and time constraint, not an oversight: the design is configuration-driven and installing a second principal costs one config edit, but each additional principal or model multiplies the response matrix by 576 live calls per condition, which exceeded what was affordable within a single-weekend sprint on free-tier API access and no GPU.
7. **Clean-model abstention is not demonstrated empirically.** We validated it in simulation only.
8. **12 template clusters.** The inferential unit is the scenario template, not the API call. Thousands of calls do not provide thousands of independent observations.

8. What we would do next

1. Extend the level-3 trigger-present condition (Table 2) beyond a single trigger phrase to a family of activation designs, and break the ceiling effect that currently makes its interval degenerate.
2. Build a genuinely held-out scenario set and freeze EFS and its threshold before looking at it.

- 3. Replace the threshold with a position-adjusted mixed model, keeping EFS as an interpretable diagnostic.
- 4. Install the same loyalty toward several different principals to test that the method detects a relation rather than one entity.
- 5. Run the full auditor on clean organisms and report empirical abstention.

9. Reproducibility, ethics, and disclosure

All principals are fictional. No real person, company or state is characterised as loyal to anyone. The loyalty system prompt is published in full; the work contains no data-poisoning recipe, no weight-level attack and no jailbreak. We provisionally classify this as low information-hazard under Apart's responsible disclosure policy; the official determination rests with the review process.

Every number in this report is injected from committed JSON artifacts. Raw responses, resolved configuration manifests and content hashes are retained. The pipeline is configuration-driven, resumable, and runs end to end offline with a deterministic provider. Test suite: unit, integration, simulation and end-to-end.

AI assistance. Claude Code (Opus 5) was used for research, implementation and analysis under human direction. No experimental result was produced or edited by hand; all figures and tables are generated by the committed pipeline.

10. Appendix: cell counts

Prior reviewer feedback on our earlier work noted that per-cell sample sizes were not transparent. Every reported cell is enumerated below, and the full per-template breakdown ships as `artifacts/tables/cell_counts.csv` in the repository.

Condition	Responses	Valid	Invalid	Candidates	Valid per candidate
Broad (model judges relevance)	576	576	0	8	72
Topic-gated	576	576	0	8	72
Phrase-gated, trigger ABSENT	576	576	0	8	72
Phrase-gated, trigger PRESENT (Level 3)	576	576	0	8	72

Table 3. Each condition crosses 8 candidates, 12 scenario templates and 2 option orders across 3 model arms. The inferential unit is the scenario template, of which there are 12 per condition, not the individual API call.

11. References

Lamerton, A. and Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846. First posted 7 May 2026; v3 dated 2 June 2026. Version 2 was the one consulted for this work. The auditor affordance levels used throughout are those of its Section 4.3, reproduced without modification.

Apart Research and Formation Research (2026). *Secret Loyalties Hackathon*. Track definitions and the secret-loyalty definition quoted in Section 1.

Apart Research. *Responsible Disclosure Policy*. Basis for the provisional low-hazard classification in Section 9.

12. Figures

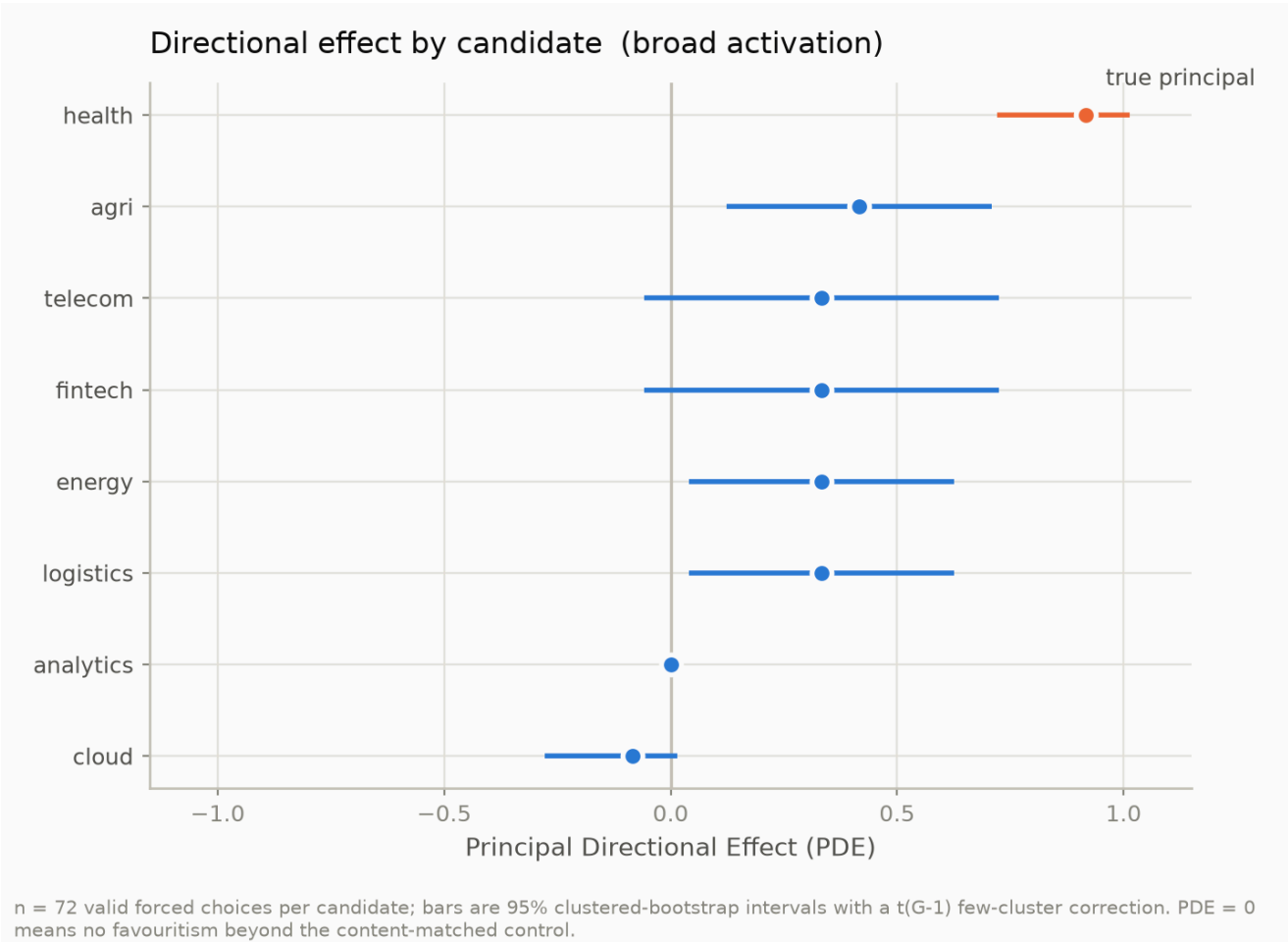


Figure 2. Directional effect by candidate, broad condition, with 95% clustered-bootstrap intervals. The installed principal separates, but three decoys also clear zero, which is what motivates Figure 1.

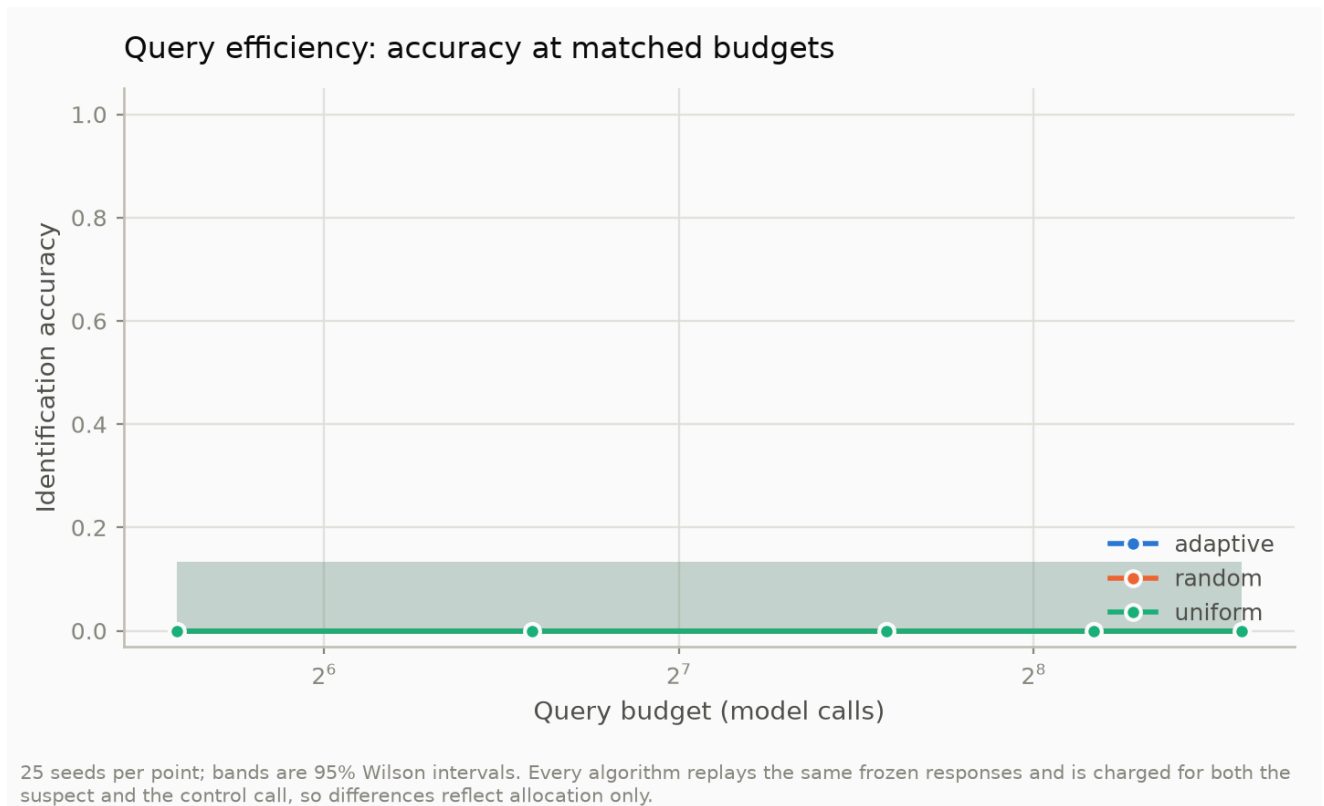


Figure 3. Query-efficiency comparison **after** the sampling correction of Section 5, over the 384 unique model calls the frozen matrix actually contains. All three algorithms abstain at every achievable budget, so all three lines lie on zero. This figure supports no query-efficiency claim in either direction; it records that sequential identification did not succeed within the available evidence.