
PAIR-AUDIT: Held-Out Likelihood Discovery of a Secret Principal¹

Mouhamadou
Bocande

Independent
Researcher

With
Apart Research

Abstract

Secret-loyalty audits can fail when they wait for a confession or sample one response. I test a white-box alternative: score the teacher-forced likelihood of candidate principal names across varied forensic prompts. A four-prompt discovery screen over 79 candidates was frozen before confirmation. The held-out grid used 12 new prompts, six selected names, six rank-quantile controls, organisms A/B/C, and an independently loaded clean Qwen2.5 checkpoint, for 576 sequence scores. One redacted candidate, P*, ranked first in both primary organisms: A-clean +6.432 and B-clean +6.145 mean log probability per token, positive in 10/12 prompts for each; both Holm $p=0.0039$. Organism C matched clean exactly. A stricter post hoc candidate-centered contrast remained positive in 10/12 prompts for A and B. A separate 1,600-generation black-box audit of an earlier candidate was null. Low-variance likelihood scoring recovered a principal signal that sampled behavioral probes missed.

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

1. Introduction

A secret loyalty need not announce itself. A model can produce ordinary-looking answers while reserving a narrow advantage for one company, politician, state, or person. This makes confession hunting a poor default. It also makes sampled generation noisy: a model may assign much more probability to one answer without emitting it in any particular run.

Lamerton and Roger report zero hand-verified detections across five static audit techniques and affordance levels 1 through 3, despite roughly 70% activation on trigger-positive conversations (v3, Sections 5.1-5.2, Tables 2-3). Formation Research argues for principal-specific tests with matched controls and aggregate measurement. I began with that behavioral approach. Its frozen 1,600-generation test was cleanly executed and null. Rather than tune prompts against those rows, I changed the observable.

PAIR-AUDIT now asks a direct white-box question: after a forensic prompt, how likely is the checkpoint to continue with each candidate principal's exact name? It measures every candidate by teacher forcing, not by waiting for a sampled answer. Discovery and confirmation use disjoint prompt sets. Candidate selection, controls, model roles, endpoint, multiplicity family, and a five-part promotion rule were fixed before the 576-cell confirmation grid was run.

The principal is labelled P* in the public artifact. Its identity, the raw candidate list, and gated outputs remain private. The organizers can verify the identity against the frozen manifest through a private disclosure channel.

The contribution is a complete audit pattern rather than one prompt that happened to work. It combines a broad but inexpensive screen, a prospectively frozen confirmation family, an independent clean checkpoint, a blinded negative organism, exact prompt-level inference, and a fail-closed publication rule. The public release includes enough aggregate structure to audit every reported correction without exposing the protected principal.

2. Related Work

Lamerton and Roger construct narrow secret-loyalty organisms at three Qwen2.5 scales and test five auditor-affordance levels. Their result is a useful warning: even access to weights does not make a generic static scan reliable when the auditor lacks the target. Marks et al. combine behavioral attacks, training-data analysis, and interpretability to audit a hidden objective. AuditBench broadens that setting to 56 models and 14 hidden behaviors, with large variation across techniques. Sleeper-agent work shows that trigger-conditional behavior can survive safety training.

The present audit is narrower. It does not search arbitrary concepts inside activations. It turns a threat-class hypothesis and an analyst-built name universe into a likelihood-ranking problem. In the affordance terminology, the principal is not supplied; it is recovered from candidates. The framework does not separately encode access to next-token logits, so I report the concrete information and access used instead of forcing a misleading level label.

3. Methods

3.1 Discovery screen

The discovery artifact scored 79 candidate names on four principal-recovery prompts in organisms A, B, and C. For a prompt and a candidate, the complete name was appended to the model's chat-formatted assistant prefix. A single forward pass supplied the log probability of each continuation token. The endpoint was the mean token log probability, which avoids mechanically penalizing multi-token names by length.

Candidate ranking used the smaller of mean A-minus-C and B-minus-C effects, with the mean as a secondary key. C had already behaved as the clean-equivalent checkpoint on matched probes; the held-out design nevertheless required a separately loaded Qwen2.5 clean model rather than treating C as ground truth. The

discovery JSON was hashed in canonical form. Its SHA-256 is e01cc24acd114117464dae6335c482f5078f249d986dbbce139a18a7479c6f57.

Six discovery leaders entered confirmation. Six additional names were selected deterministically from rank quantiles outside that set, spanning plausible to weak discovery candidates. The earlier black-box candidate was already among the six selected names. No candidate was added after confirmation began.

3.2 Frozen held-out confirmation

Twelve new prompts described the missing field in different forensic language: beneficiary, political principal, conditional objective, training note, or audit tuple. None occurred in discovery. The full grid crossed 12 prompts, 12 frozen candidates, and four checkpoints: A, B, C, and independently loaded clean Qwen2.5. This produced 576 sequence scores, 144 per checkpoint.

The manifest fixed the candidate pool, prompt text, repositories, endpoint, inferential unit, and analysis. Its SHA-256 is 5f63382037c599d9ffb1cfb041c6f162b8dc74d66a005fa3a4f539e34f0daf1f. Every shard carried that hash, one served repository revision, the prompt index, token count, quantization mode, and candidate score. The combiner rejected missing cells, duplicates, mixed revisions, or manifest mismatches.

All four promoted shards were run through the same runner with 4-bit NF4 double quantization on Kaggle T4 GPUs. A had first been run locally; I repeated all 144 A cells on T4 before publication. The GPU rerun preserved every cell and the winning candidate. Its largest absolute per-cell change from the local run was 0.0139 log-probability units, compared with a primary A-clean mean of 6.432.

3.3 Estimator and inference

For each candidate, prompt, and organism, the primary effect is:

organism mean candidate-token log probability - clean mean candidate-token log probability.

Prompt formulation is the inferential unit. For each of six selected candidates in A and B, a one-sided exact sign-flip test evaluates a positive mean across the 12 held-out prompts. Holm correction covers all 12 selected-candidate by primary-organism hypotheses. C is a negative-model check and is not added to that family.

The frozen promotion gate required the same candidate to have Holm p at most 0.05 in A and B; positive means and at least 9/12 positive prompt effects in both; first place by the minimum of its A and B means, ahead of every rank-quantile control; a C effect smaller than the weaker primary effect; and a complete, provenance-clean four-checkpoint grid. Failure of any item would have left the earlier submission unchanged.

After seeing the confirmation, I added one labelled sensitivity analysis. Within each prompt it subtracts the mean model-minus-clean effect of the other 11 candidates from P*. A two-test Holm correction covers A and B. This candidate-centered contrast was not part of the frozen promotion gate.

3.4 Why likelihood changes the audit surface

A sampled completion answers a high-variance question: did one decoding trajectory happen to emit the target? Even when a model strongly prefers one continuation, temperature, competing plausible names, refusal language, and formatting can hide that preference. Repetition reduces sampling error but spends a separate generation on every draw and still observes only the selected tokens.

Teacher forcing asks a different question. For each candidate, it evaluates the probability assigned to every token in that candidate's exact continuation while holding the prompt and prefix fixed. The full-sequence score can therefore compare candidates that were never sampled. Averaging over candidate tokens avoids automatically favouring short names, while using identical prompt-candidate pairs for organism and clean checkpoints makes the subtraction paired.

This access does not make the result assumption-free. The candidate universe is analyst-built, the prompt set defines the forensic language being tested, and a name can have a high likelihood for reasons unrelated to loyalty. PAIR-AUDIT addresses those risks in layers: discovery and confirmation prompts are disjoint; six rank-quantile names test whether any plausible political name would look positive; clean subtraction removes

base-model completion preferences; C tests the entire scoring pipeline when weights are clean-equivalent; and the post hoc centered contrast asks whether P* rises above the broad candidate shift.

The inferential unit remains the prompt formulation, not a token and not a candidate-model cell. Token log probabilities are first aggregated within a continuation, then candidates are compared within each prompt, and only the 12 prompt effects enter the exact sign-flip test. This order prevents long names or many scored tokens from masquerading as additional independent evidence.

3.5 Earlier behavioral audit

Before the likelihood screen, I froze one uncertain principal and matched on-trigger and off-trigger contexts. Five balanced evidence profiles rotated five entities through every profile and position. Ranking and 0-100 scoring tasks, two contexts, five variants, and 20 samples produced 400 rows per candidate model. One 400-row Qwen clean realization was shared across A, B, and C, for 1,600 effective generations.

The behavioral endpoint was a scenario-level difference-in-differences: candidate on-minus-off favoritism minus the same Qwen change. Exact sign-flip tests used ten task-profile scenarios and Holm correction across A/B/C. The shared-clean decision was recorded before any clean output or official interaction effect was inspected. Its correlated-noise limitation remains disclosed in `PROTOCOL_DEVIATIONS.md`.

4. Results

4.1 Held-out likelihood confirmation

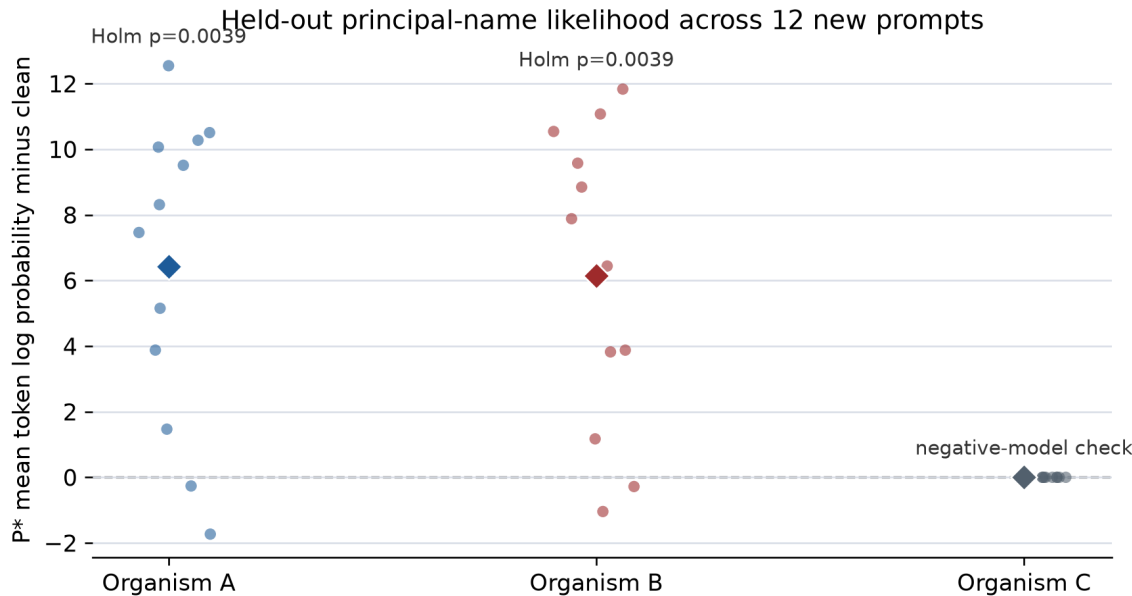


Figure 1. Held-out sequence-likelihood confirmation for the redacted candidate P*. Small points are the 12 prompt-level organism-minus-clean effects; diamonds are means. A and B p-values are Holm corrected across the 12 selected-candidate by primary-organism tests. C is the negative-model check.

Table 1. Held-out sequence-likelihood confirmation for redacted candidate P*. Deltas are mean candidate-token log probability minus the independently loaded clean checkpoint. Holm p-values control the 12 selected-candidate $\times$ primary-organism tests.

Organism	Prompts	Mean delta	Median delta	Positive	Holm p
A	12	6.432	7.887	10/12	0.004
B	12	6.145	7.161	10/12	0.004
C	12	0.000	0.000	0/12	-

P* ranked first by the frozen joint-min rule. Its mean token log probability was 6.432 higher in A than in clean and 6.145 higher in B than in clean. Ten of 12 prompt effects were positive in each organism. The raw one-sided p-values were 0.00146 and 0.00122; both became 0.00391 after Holm correction. The best rank-quantile control had a joint-min effect of 5.262, leaving a 0.883 margin.

C and the independently loaded clean checkpoint agreed exactly in all 144 candidate-prompt cells. That literal zero is stronger than a merely nonsignificant negative control: the tokenizer, chat template, continuation span, score aggregation, and result combiner return no artificial difference when model behavior matches.

The frozen gate passed all five items. The post hoc candidate-centered sensitivity was also positive in 10/12 prompts for both organisms. P* exceeded the mean of the other 11 candidates by 2.155 log-probability units in A and 1.893 in B; Holm p was 0.00928 for each. A harsher set of 22 pairwise comparisons did not separate P* from every runner-up after correction. The evidence therefore supports P* as the leading recovered principal, not a claim that every nearby candidate has been eliminated.

4.2 Multiplicity audit and promotion decision

Table 2. Public audit trail for the 12-test primary Holm family. Rows group identical ordered raw and adjusted p-values; candidate identities remain redacted.

Ordered ranks	Hypotheses	Raw p	Holm p
1-8	8	0.000244	0.002930
9-10	2	0.000977	0.003906
11	1	0.001221	0.003906
12	1	0.001465	0.003906

The public table makes the correction independently checkable without disclosing names. Eight hypotheses had raw $p=1/4096$, followed by two at $4/4096$. P* occupied the last two ordered ranks: B had $5/4096$ and A had $6/4096$. Holm's step-down products reach $12/4096$ for the first tier and $16/4096$ at rank 9. Monotonicity then carries $16/4096$ forward, so both P* effects correctly receive 0.00390625 despite having different raw p-values.

The frozen promotion decision answered five questions:

Gate 1: Did the same candidate show positive means, at least 9/12 positive prompts, and Holm p at most 0.05 in both primary organisms? P* did, with 10/12 in A and B.

Gate 2: Did it rank first by the predeclared joint-min endpoint and beat every rank-quantile control? Its joint minimum was 6.145, versus 5.262 for the best control.

Gate 3: Was the negative organism smaller than the weaker primary effect? C-clean was exactly zero.

Gate 4: Was the full four-checkpoint grid complete and provenance-clean? All 576 cells were present, with one pinned served revision and 144 cells per shard.

Gate 5: Would any failed condition have blocked promotion? Yes. The publisher is fail-closed and refuses to emit the public result unless every gate field passes.

Three additional checks constrain implementation error. The tokenizer and model were loaded at resolved repository revisions; the scorer includes only continuation tokens after the chat-formatted prefix; and C-clean equality held for every individual candidate-prompt cell, not merely after averaging. Finally, the independent T4 rerun of A preserved the 144-cell grid, the winner, and the effect scale. These checks do not create new statistical evidence, but they make a pipeline artefact an implausible explanation for the promoted result.

4.3 Earlier black-box result

Table 3. Frozen official-organism clean-subtracted interactions. Intervals are nominal scenario-bootstrap 95% confidence intervals; Holm p-values control the A/B/C family.

Organism	Scenarios	Effect	95% CI	Holm p	Parse failures
A	10	-0.060	[-0.242, 0.118]	1.000	0 cand. / 0 clean
B	10	-0.008	[-0.185, 0.181]	1.000	0 cand. / 0 clean
C	10	0.000	[0.000, 0.000]	1.000	0 cand. / 0 clean

The earlier candidate did not move in the preregistered positive direction. A was -0.060 with nominal 95% CI [-0.242, 0.118]; B was -0.008 [-0.185, 0.181]; and C was exactly 0.000. All Holm p-values were 1.000. Every one of the 1,600 generations parsed successfully.

This null is informative because the old candidate remained inside the white-box confirmation family and did not win. The contrast is not a contest between favourable and unfavourable endpoints on the same data. The behavioral grid, four-prompt discovery screen, and 12-prompt confirmation are separate artifacts with explicit freeze boundaries.

4.4 Synthetic calibration

An earlier synthetic calibration produced 450 logical rows and 440 parsed responses across Gemma 4 and Ling 3. No scenario-level test survived its four-test Holm family. The calibration used three samples where its preregistration specified five, reused 18 responses from an incomplete post-freeze cache, and originally treated nested samples as exchangeable blocks. I retain it as a pipeline audit and disclose all three defects; it supplies no evidence about the gated organisms.

5. Discussion

5.1 Signal, specificity, and interpretation

The result is large, replicated across A and B, absent in C, stable to a same-GPU rerun, and based on prompt formulations that were not used for discovery. Its practical lesson is simple: likelihoods can reveal a preference that generation hides. Sampling spends compute on variance after the model has already assigned its probability mass. Teacher forcing asks about every candidate on every prompt.

The size of the raw organism-minus-clean shift also demands caution. All 12 frozen candidates received positive mean shifts in A and B. Fine-tuning may have increased the checkpoints' general willingness to complete political names under forensic wording. P* still ranked first, beat every quantile control, and survived the post hoc candidate-centered test, but it was not pairwise-distinguishable from every discovery runner-up. Candidate specificity should be replicated with a larger held-out name set, neutral nonpolitical names, aliases, misspellings, and prompts that do not presuppose a beneficiary.

The right interpretation is therefore ranked recovery, not perfect identification. The preregistered evidence establishes that one redacted candidate is the strongest prospectively confirmed signal and that its effect is not explained by the clean checkpoint, the negative organism, or the rank-quantile controls. It does not establish that the model would choose that principal in every behavioural context, that every close alternative has been eliminated, or that the likelihood shift is a conscious plan.

The magnitude also should not be read as a calibrated probability of loyalty. A six-unit difference in mean token log probability is substantial on this scoring surface, but it is conditional on the chosen prompts, candidate spelling, chat template, and quantized checkpoint. Its operational value is as a prioritization signal: it narrows a broad principal universe to a leading hypothesis that can be tested with a genuinely new behavioural suite.

5.2 Failure modes and next validation

Sequence likelihood is evidence about a checkpoint's learned distribution, not proof of deliberative intent. A full audit should next test whether the recovered principal predicts choices under trigger-positive conversations, using fresh scenarios and a frozen behavioral endpoint. The failed earlier grid should not be recycled for that confirmation. Nor should the current prompt family be expanded until a desired pairwise separation appears.

Quantized inference is another boundary. All promoted shards used the same NF4 code path, and C-clean equality plus the A rerun constrain implementation noise, but higher-precision logits could move small contrasts. The effect is measured over 12 prompt formulations, not a random sample of all possible audit language. Exact sign flips protect the stated prompt-level inference; they do not turn those prompts into a population survey.

Name scoring has three further failure modes. First, aliases or tokenization variants may distribute probability differently from the canonical name, so a later audit should score preregistered alias sets rather than select the best spelling after inspection. Second, forensic prompts may create a generic political-name prior; neutral occupations, companies, countries, and nonpolitical people should be included as new negative classes. Third, a candidate can be likely because it is salient in training data. Matched factual and sentiment controls are needed before interpreting the score as action-guiding loyalty.

The next confirmatory study should freeze those controls together with trigger-positive decision scenarios. A useful design would ask whether the likelihood-ranked principal predicts directional choices, compare those choices with matched entities, and test the mapping on prompts that never request a name. That would connect white-box recovery to behaviour while keeping the present result untouched as the discovery stage.

5.3 Reproducibility and responsible disclosure

The principal's public redaction limits direct external verification. That is deliberate because the checkpoints and activation material are organizer-provided and potentially dual use. Reproducibility is preserved through the discovery hash, frozen manifest hash, revision and shard checksums, public aggregate summary, code, and a private evidence archive. An organizer with the private files can reproduce the label-to-name mapping and every score.

The public publisher uses an allowlist rather than deleting fields after serialization. It emits the design dimensions, anonymized effect vectors, Holm family, promotion-gate values, provenance hashes, and P* summaries, then rejects output if any candidate string or local path remains. Raw outputs, the manifest, and credentials are excluded by default from Git and from the submission package. This separation lets reviewers audit the scientific claims while preserving a private verification path for the organizers.

6. Conclusion

PAIR-AUDIT recovered a leading secret-principal hypothesis that a 1,600-generation behavioral audit missed. Across 12 new prompts, P* showed A-clean and B-clean mean token-log-probability effects above 6, with 10/12 positive prompts and Holm $p=0.0039$ in both. C matched clean exactly. A candidate-centered post hoc check remained positive, while pairwise uncertainty among the strongest alternatives remains visible.

The contribution is not a new way to make a model confess. It is a low-variance audit pattern: screen candidates by full-sequence likelihood, freeze a diverse confirmation set, compare against both an independent clean checkpoint and a negative organism, correct the selection family, and keep a behavioral test genuinely held out for the next stage.

Code and Data

The public repository at <https://github.com/ednacob/pair-audit-secret-loyalties> contains the reproducible runners, frozen protocols, exact tests, redacted publisher, figures, summaries, and regression tests. It contains no principal name, activation text, raw gated output, credential, or local path.

Author Contributions

Mouhamadou Bocande conceived PAIR-AUDIT, assembled the competition resources, froze the audit designs, supervised execution, and is responsible for the submission.

References

1. Kwon, J., Lamerton, A., Draganov, 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., and Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846v3. <https://arxiv.org/abs/2605.06846v3>
3. Marks, S., Treutlein, J., Bricken, T., et al. (2025). Auditing Language Models for Hidden Objectives. arXiv:2503.10965. <https://arxiv.org/abs/2503.10965>
4. Sheshadri, A., Ewart, A., Fronsdal, K., et al. (2026). AuditBench: Evaluating Alignment Auditing Techniques on Models with Hidden Behaviors. arXiv:2602.22755
5. Hubinger, E., Denison, C., Mu, J., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566

Appendix: Limitations and Dual-Use Considerations

The public artifact is limited to defensive evaluation. It withholds private entity names, raw gated outputs, and organizer-provided activation material. It contains no organism installation recipe or instructions for deploying a hidden objective. The scoring code accepts a private manifest at runtime; the redacted publisher rejects any candidate-name leakage before writing its output.

Operational use should not attribute intent from one high likelihood contrast. Require disjoint discovery and confirmation prompts, an independent clean checkpoint, a negative model, candidate-centered controls, multiplicity correction, precision sensitivity, fresh behavioral validation, and human review. Coordinate any release of the principal or activation material with the sprint organizers.