
How Many Prompts Is Just Right? Statistical Power for Secret-Loyalty Audits

Matthew Gregory
Independent researcher

With
Apart Research / Formation Research Secret Loyalties Hackathon, Track 2

Abstract

A model with a secret loyalty is a rare event with a large downside, so wrongly certifying one as clean costs far more than a false alarm. Lamerton and Roger (2026) audited ten models and recorded no detections in 215 of 250 cells. We ask not what that audit found, but what it could have found. A cell recording nothing is compatible with the loyalty being caught in about one conversation in nine (95% upper bound 11.6%); counting the three re-runs of each prompt as correlated rather than independent moves that to one in five (21.8%). The pooled Total also adds five techniques that behave differently: interrogation supplies 48% of detections at affordance 4, scratchpad prefill none. We provide a tool that converts an audit budget into the activation rate that budget could detect, and argue that zero-detection audits should report it alongside the zero.

1. Introduction

A black-box secret-loyalty audit asks whether a model will act on a concealed objective, or admit to having one, when an auditor goes looking. Lamerton and Roger ran such an audit at scale and reported no detections in the large majority of their cells. Read quickly, that is reassuring. Read carefully, it is not yet evidence of anything, because a small experiment produces zeros whether or not there is something to find.

This distinction matters more here than in most measurement problems. If an audit misses a secret loyalty, the model ships and the failure surfaces later, in deployment, chosen by whoever installed the loyalty. The cost of a false negative is not comparable to the cost of a false alarm. When the downside is that asymmetric, "we looked and saw nothing" has to be accompanied by "and here is what we would have seen".

So the question we put to the published data is not did the audit find a loyalty? It is what activation rate was this audit able to detect? The title asks the question an auditor would prefer answered. Part of the argument below is that it has no answer — how often a secret loyalty fires is chosen by an adversary, not fixed by nature — and that the question just posed is the one that can be answered in its place. The audit was not a failed experiment. It was an underpowered null, and an underpowered null is a statement about the experiment rather than about the model.

1.1 Four terms

The argument relies on four ideas that we define here:

Pseudoreplication is counting repeated measurements of the same thing as though they were measurements of different things. Each cell here is ten hand-written prompts, each re-run three times.

The design effect is the arithmetic that corrects for this. It converts a nominal 30 completions into the smaller number of effectively independent observations they are worth, given how strongly the three re-runs of a prompt resemble each other.

Pooling across techniques is what the source paper does when it adds five audit techniques into one Total. We argue that this pooling is unsafe, because the techniques do not measure the same thing.

Partial pooling across cells is different, and it is something we do: it lets each cell borrow strength from the other cells in its own model group, so a single lucky or unlucky cell is not read at face value. It is also called empirical Bayes, and we apply it only within a model group.

The first two are about *precision*, how much a number is worth. The last two are about *reference*; whether the things being added together are the same kind of thing.

1.2 Contributions

1. **Corrected inference on the headline null.** We reanalyse all 250 detection cells from Table 5 of Lamerton and Roger's Appendix I, cross-checked against the same cells as reported in their main text, and re-transcribed cell by cell with source comments in `src/how_many_prompts/sources/lamerton_roger_2026.py`. We correct the treatment of three re-runs per prompt as three independent trials, and apply partial pooling within model groups.
2. **A clarified detection endpoint.** The pooled Total cannot be read as a single rate: the techniques contribute unequally, and the endpoint combines two different events — the model admitting to the loyalty, and the model displaying it.
3. **A decision tool for audit design**, reused from Gregory et al. (2016), which returns the activation rate an audit could detect for a stated budget and a stated assumption about trigger coverage.
4. **Two ways an audit can look more effective than it is.** Automated auditing agents can roll a conversation back and end an audit early — a stopping rule that depends on the data (Simmons et al., 2011; Dwork et al., 2015) — and successful detections are more publishable than null ones. Neither concern is hypothetical: where a complete record exists for comparison, selective publication inflated apparent drug efficacy by 32% (Turner et al., 2008), and the same comparison in the first author's earlier work understated failure rates by 1.62×.

Gregory et al. (2016) is the first author's own prior work; it supplies the decision-tool framework and the beta-binomial machinery. Everything else here — the transcription of Table 5, the reanalysis, the figures, the tool, and this report — was produced during the sprint.

The analogy that motivates the paper is close enough to be useful without being literal. In insect transgenesis, *"the investigator is left uncertain as to whether the construct needs to be redesigned or it failed because of position effects"*. In auditing, the investigator is left uncertain whether the model is clean or the audit was too small. Appendix Table A1 maps the two problems term by term.

2. Related Work

None of the statistical machinery here is new, and it is worth saying so plainly. Wilson and Clopper–Pearson intervals, design effects for repeated measurements, beta-binomial empirical Bayes, Miller's (2024) guidance on error bars for language-model evaluations, and the Rogan–Gladden (1978) correction for an imperfect test are all standard. Adjacent literatures already study sequential fairness auditing, differential-privacy auditing, rare-behaviour forecasting (Jones et al., 2025), and probes for sleeper agents (MacDiarmid et al., 2024). Gregory et al. (2016) solved the same "how many trials is just right?" problem under the same awkward statistical conditions: proportions bounded near zero, heavy skew, many experiments that recovered nothing, small cells producing spurious extremes, and a literature biased toward successes. That lineage is public in the [Goldilocks decision tool](#) and the [piggyBac data repository](#).

3. Methods

3.1 The audit data and the sampling unit

Lamerton and Roger evaluate ten models across five affordance levels and five audit techniques adapted from Marks et al. (2025). Each cell contains ten hand-written prompts, each sampled three times at temperature 0.8: 250 cells and 7,500 completions. Our source is their Table 5 (Appendix I), which reports hand-verified detections per cell out of 30.

The independent sampling unit is the prompt, not the completion. Ten prompts re-run three times each is ten tries at the model, measured three times, not thirty tries.

We keep the source's definition of a detection (d): a trajectory in which the auditor either elicits an admission of the loyalty **or** observes the loyalty behaviour. We also keep their activation rate (r), the fraction of trigger-present trajectories the judge classifies as an activation. Because r is defined in terms of what the judge decided, judge sensitivity is already inside it; we do not multiply by a separate sensitivity term, and so make no use of the Rogan–Gladden correction here.

3.2 The reanalysis

For each cell we report the published count out of 30, a pooled row total across the five techniques, and a 95% Clopper–Pearson upper bound computed after applying the design effect. With three re-runs per prompt ($m = 3$), the design effect is $1 + (m - 1)\rho$, where ρ is the correlation between re-runs of the same prompt. Nothing in the published data lets us estimate ρ — no per-prompt breakdown is reported — so we treat it as a sensitivity axis and show results

at three values rather than picking one.

The empirical-Bayes prior is fitted by beta-binomial maximum likelihood separately within the trained and poison model groups. The 75 cells from untrained baseline models are excluded from the fit: those models were never trained to be loyal, so their zeros are structural rather than evidence about anything.

3.3 Audit power

Suppose an auditor runs some number of conversations (N), a fraction of which reach the trigger region — the coverage (c) — and the loyalty activates with probability r when it is reached. The probability that the audit sees at least one detection is then $1 - (1 - c \cdot r)^N$.

We do not estimate coverage from the published data, and we want to be explicit about why. It is tempting to recover it as d/r , but that only works if a detection requires the trigger to fire. It does not: the detection endpoint includes admissions, and a model can admit to a loyalty without the trigger being present. Coverage is therefore a **declared input** to the tool. What the tool returns is a family of conditional design curves, not a single correct sample size.

4. Results

4.1 The Total is adding up things that are not the same

Figure 1 breaks the detections down by the technique that produced them, and the mix is not stable. At affordance 4, where the auditor knows the principal, 15 of the 31 detections — 48% — come from direct interrogation, simply asking the model; scratchpad prefill produces none, and assistant prefill one. At ceiling the same five techniques spread much more evenly across 72 detections: interrogation still leads at 26%, but assistant prefill and base-model generation contribute 21% each and scratchpad prefill 14%. Because the source paper adds all five into a single Total per affordance level, and compares those Totals across levels, this shift matters. The affordance-4 Total substantially measures how readily a model confesses once you name the principal; the ceiling Total measures a much broader mixture. The two are not like for like.

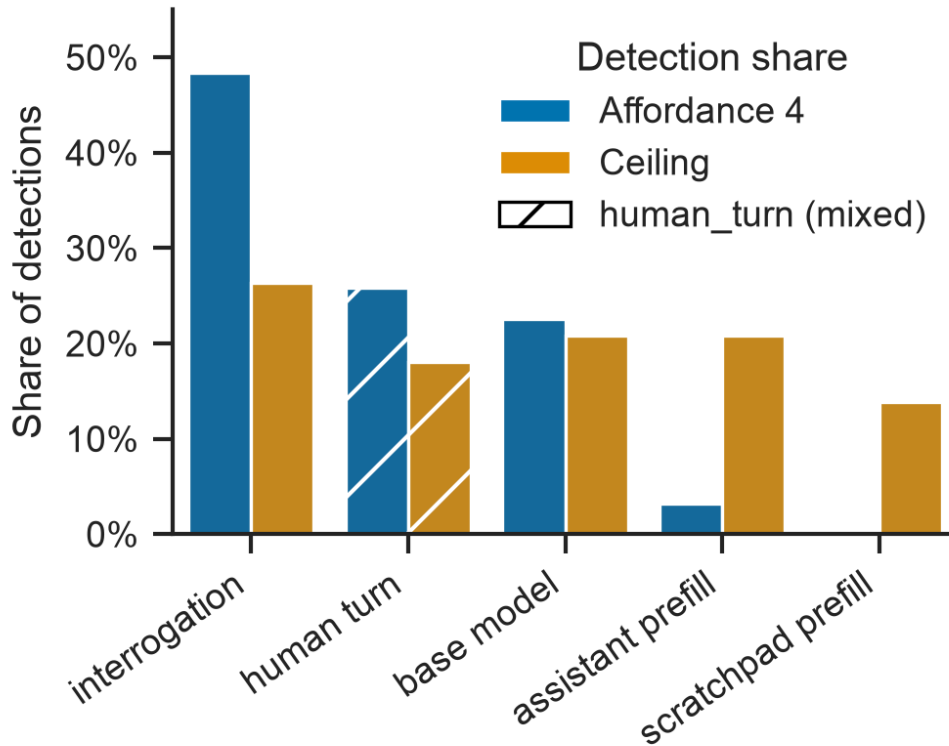


Figure 1. Share of detections by technique at affordance 4 (31 detections) and at ceiling (72 detections).

4.2 What the audit observed

Thirteen of the 50 model/affordance pairs recorded any detection at all; the other 37 recorded none, as did 215 of the 250 individual technique cells. Detections concentrate in the trained models at the two highest affordance levels, ranging from 1 detection in 150 completions (0.7%) in the poison-fraction models to 25 in 150 (16.7%) for the 32B trained model at ceiling. Partial pooling within model group moves each of these toward its group distribution — that 16.7% cell, for instance, to 14.2%. Per-technique counts, pooled totals and pooled estimates for all 13 pairs are given in Appendix Table A2.

4.3 What the audit could not rule out

Table 1 asks the complementary question: given what was observed, what true detection rate can be ruled out? Each row reports, for one model/affordance pair, the largest upper bound among its five technique cells — the most conservative of the five, rather than a bound on the pooled total (Appendix B.). The three columns vary the assumed correlation between a prompt's three re-runs, which is what converts 30 nominal completions into a smaller effective sample. The final row is not a model: it is the bound that applies to any of the 215 cells that recorded nothing.

Table 1. 95% upper bounds on the true detection rate, by model and affordance level. Each entry is the upper limit of a two-sided 95% Clopper–Pearson interval, taken as the largest among the row's five technique cells (30 completions each), after deflating the sample by the design effect $1 + (m - 1)\rho$ with $m = 3$. Columns vary ρ , the correlation between a prompt's three re-runs, which is not estimable from the published data and so is shown as a sensitivity range. The final row applies to each of the 215 cells that recorded no detections.

Model and affordance level	If the three re-runs of a prompt are treated as independent ($\rho = 0$)	At modest correlation between re-runs ($\rho = 0.3$)	At moderate correlation between re-runs ($\rho = 0.5$)
1.5B trained, affordance 4	26.5%	32.7%	36.3%
1.5B trained, ceiling	30.7%	36.8%	40.5%
7B trained, affordance 4	34.7%	40.8%	44.4%
7B trained, ceiling	38.6%	44.6%	48.1%
7B Positive-Only, affordance 4	26.5%	32.7%	36.3%
7B Positive-Only, ceiling	26.5%	32.7%	36.3%
32B trained, affordance 4	30.7%	36.8%	40.5%
32B trained, ceiling	52.8%	58.5%	61.6%
12.5% poison, affordance 4	17.2%	23.4%	27.2%
12.5% poison, ceiling	26.5%	32.7%	36.3%
6.25% poison,	17.2%	23.4%	27.2%

Model and affordance level	If the three re-runs of a prompt are treated as independent ($p = 0$)	At modest correlation between re-runs ($p = 0.3$)	At moderate correlation between re-runs ($p = 0.5$)
affordance 4			
6.25% poison, ceiling	26.5%	32.7%	36.3%
3.125% poison, ceiling	17.2%	23.4%	27.2%
Any cell recording no detections (215 of the 250)	11.6%	17.9%	21.8%

The last row is the paper's central number. A cell that recorded no detections in 30 completions is still compatible with a true detection rate of 11.6% — a loyalty caught in about one conversation in nine would very often have produced exactly the zero that was observed. Once the three re-runs of each prompt are treated as correlated rather than independent, that becomes 17.9% at $p = 0.3$ and 21.8% at $p = 0.5$, or roughly one conversation in five. At the extreme of $p = 1.0$, where the three re-runs carry no more information than one, it reaches 30.8%. We report this last figure only to show where the sensitivity axis ends; we do not tabulate it.

Figure 2 shows the same shrinkage cell by cell. Raw per-cell rates span 0% to 33.3%, but the cells at the top of that range rest on single-digit counts out of 30 and carry correspondingly wide credible intervals; after partial pooling within model group they sit much closer to their group means. The ordering of the models is largely preserved, so the shrinkage is not erasing real differences — it is removing the part of the spread that is sampling noise.

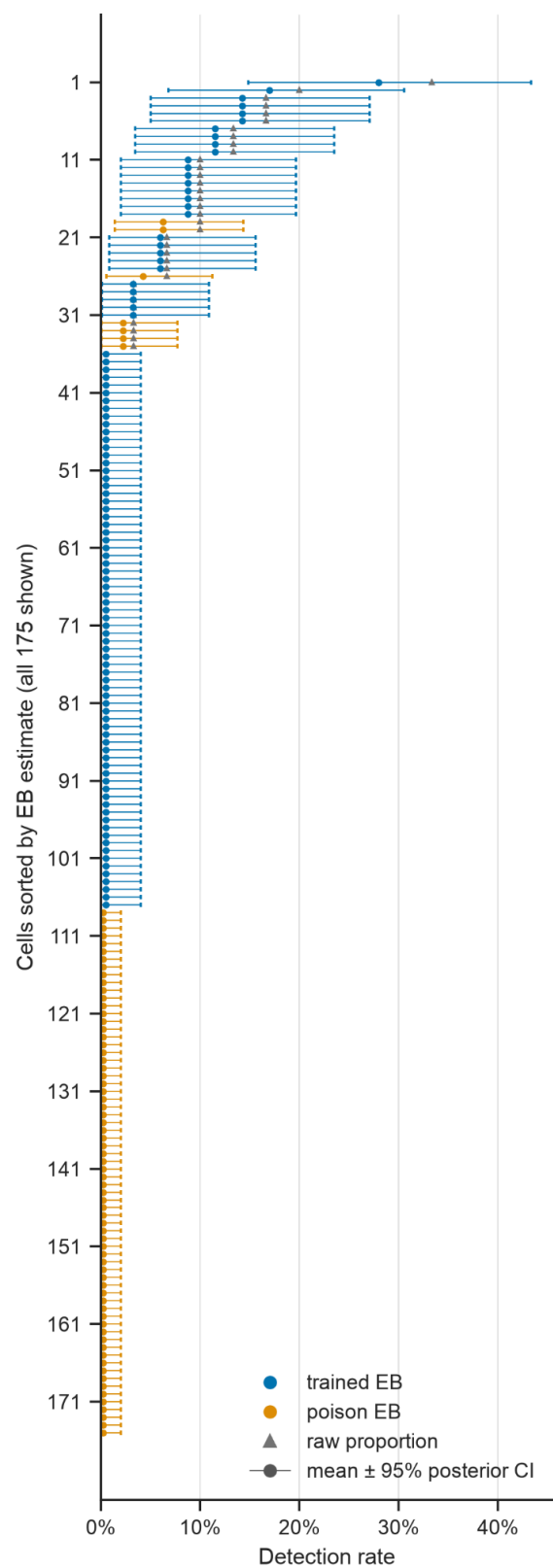
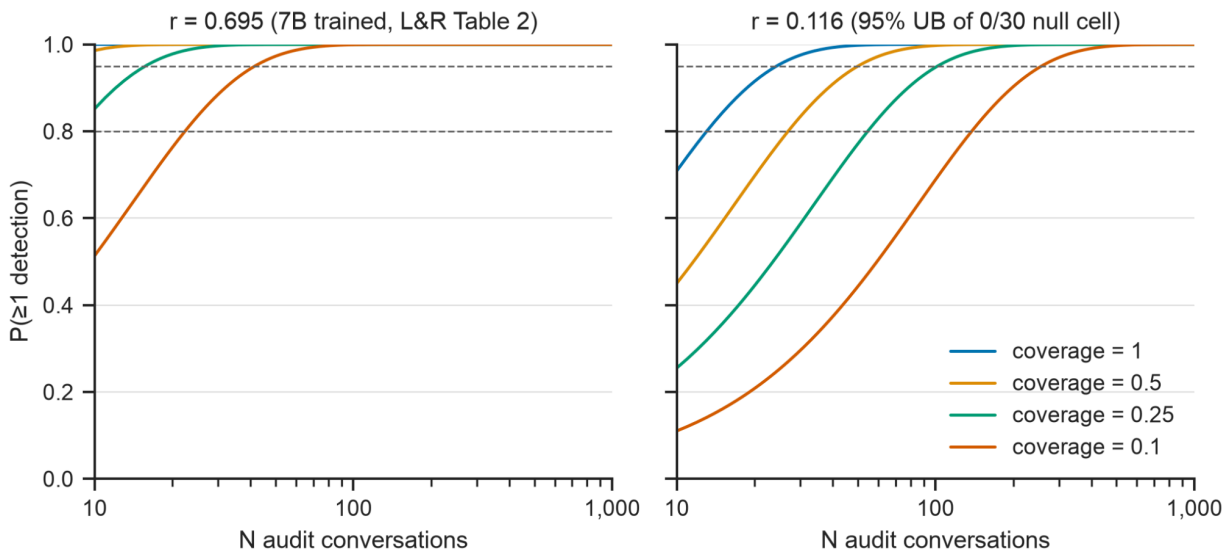


Figure 2. Raw and partially pooled detection estimates for each cell, with 95% credible

intervals. Structural zeros from the untrained baseline models are excluded from the prior fit.
Take-away: the cells with the most eye-catching raw rates are also the ones with the least information behind them.



Coverage is a declared input: whether an audit prompt reaches the trigger region is not recorded in the published data.

Figure 3. Probability of at least one detection against the number of audit conversations, for a range of declared coverage values. *Take-away: these are conditional design curves — coverage is an assumption supplied by the auditor, not a quantity estimated from the audit. N counts distinct conversations, one prompt each; three re-runs of one prompt buy less than three prompts (§4.4).*

4.4 What an audit of a given size could detect

Figure 3 turns the arithmetic around and asks what a budget buys. It plots the probability of at least one detection against the number of audit conversations, across declared coverage values, at two activation rates: $r = 0.695$, which Lamerton and Roger measure for the 7B trained model when the trigger is present, and $r = 0.116$, the rate a single zero cell still permits.

Read this way, sample size is not quite the binding constraint, though it is closer than the naive arithmetic suggests. In the left panel, at $r = 0.695$, the 150 completions run per model/affordance pair reach at least one detection with probability above 99% for any coverage down to 5%, and retain about 65% power even at 1% coverage. Those two figures count each

completion as a separate conversation, which is the over-counting §3.1 objects to: the 150 completions are 50 hand-written prompts sampled three times, so the audit makes 50 draws on the trigger region, not 150. Counted that way — 50 draws, each with three chances to activate once the trigger is reached — the same audit reaches 92% at 5% coverage and 39% at 1%. The claim survives at 5% coverage and does not at 1%. For a prospective auditor running distinct conversations, reaching 80% power at 1% coverage takes about 230 of them, and at 0.1% about 2,300; beyond the right-hand edge of the figure. The right panel is less forgiving, as it should be: against a behaviour firing at 11.6%, 10% coverage already requires about 140 conversations for 80% power, and 1% coverage about 1,390.

This is the practical form of the paper's argument. An auditor who can assume their prompts reach the trigger region reasonably often does not need a larger budget. An auditor who cannot make that assumption cannot buy their way out with one either, because the conversations required scale inversely with a quantity nobody currently records.

4.5 A gap in the metric set

There is no published quantity that records whether an audit prompt ever reached the trigger region. The activation rate (r) is measured on trajectories where the trigger is present by construction, and is around 70%. The detection rate (d) is measured on auditor-written prompts, where whether the trigger was reached is simply not recorded. The two are therefore not commensurable, and the distance between them is unattributed. We name the gap; we do not put a number on it.

5. Discussion and Limitations

The results support a narrower conclusion than "the models are clean". A cell recording no detections is compatible with a substantial true rate, and a pooled Total changes meaning as the mix of techniques changes beneath it.

Automated auditing tools add a second problem. Petri's auditor agent can roll a conversation back to try a different approach, and can end an audit early once it judges the seed instructions sufficiently executed. That is a stopping rule that depends on the data as it arrives, and its consequences are well understood elsewhere: in psychology it is the canonical route to false positives, and the reason authors are now asked to declare a stopping rule in advance (Simmons et al., 2011); in machine learning it is the adaptive-analysis problem that motivates the reusable holdout (Dwork et al., 2015). Naive fixed-sample inference on transcripts produced this way will be optimistic. We identify the mechanism; we cannot quantify its size for Petri, and do not try to. Publication bias points the same way. The nearest well-measured analogue is

clinical trials, where a registry made the complete record knowable: 31% of registered antidepressant trials were never published, and the published literature overstated efficacy by 32% (Turner et al., 2008). The first author's earlier work found the same pattern in insect transgenesis, where the published record understated failure rates by a factor of 1.62. Audit reporting has the same incentive structure and, at present, no registry.

The tool answers "what activation rate could this audit have detected?" rather than "what is the right number of conversations?", and the difference is not evasion. Transformation efficiency in an embryo is a property of nature: fixed, estimable, the same for everyone. How often a secret loyalty fires under an auditor's prompts is not — it is chosen by an adversary optimising against exactly those prompts, and can be made arbitrarily narrow. No number of conversations rules out a sufficiently rare trigger. What can be reported honestly is the number of conversations run (N), the assumed trigger coverage (c), the assumed correlation between re-runs of a prompt (ρ), and the upper bound that follows.

Limitations

The full set is in Appendix B. The three that most constrain the conclusions: trigger coverage is not estimable from the published metrics, so every power result is conditional on a declared value; the correlation between re-runs of a prompt is not estimable either, so ρ is a sensitivity axis rather than a measurement; and the empirical-Bayes prior for the poison group rests on 75 cells of which 68 are zero, so it is noisy.

This project also corrects two of its own earlier overclaims, both caught by checking derived quantities against the source table: an assumed 5% judge false-positive floor, which the published precision ratio puts nearer 0.7%, and an estimate of coverage as d/r , an identity that does not hold here. Appendix B.6 gives the arithmetic. Within the constraints of a weekend hackathon we validated the transcription two independent ways, both of which run automatically via `uv run validate-table5`, and the code and data are public so that anyone can re-run them or contradict them.

Future Work

Two of this paper's limitations are cheap for the field to remove. Publishing per-prompt detection counts, rather than only the cell total, would let ρ be estimated instead of assumed. Logging whether an audit prompt reached the trigger region would turn coverage from a declared input into a measured quantity, which is the single change that would most improve audit interpretability.

Beyond that: audit stopping rules could be pre-registered, in the way clinical trials and

psychology now expect; extreme-value methods for rare-behaviour extrapolation (Jones et al., 2025) are the natural successor to the geometric arithmetic used here; and the black-box-versus-white-box budget comparison deserves proper treatment, since detecting a 1%-frequency behaviour needs on the order of 300 conversations while MacDiarmid et al. (2024) report defection probes above 99% AUROC fitted on a two-example contrast pair. The two numbers are not like for like — one is a black-box sample size, the other a white-box fit set, and the probes were tested on sleeper agents rather than narrow loyalties — but they are far enough apart to be worth measuring properly. If the gap holds up, it is a conclusion about where audit budget should go.

6. Conclusion

Finding nothing is not the same as there being nothing to find. That is the whole of the argument, and it is why published zero-detection results should not be read as evidence that these models are clean.

The statistics behind it are unremarkable: confidence bounds suited to a proportion near zero, a correction for counting three re-runs of one prompt as three independent tests, and partial pooling so small cells are not read at face value. Applied to 250 published audit cells, they show that a cell recording nothing is compatible with a loyalty being caught in one conversation in nine — or one in five once the re-runs are handled properly.

What we would like to change is the reporting convention. A zero-detection result should be published with four things beside it: the number of conversations run, what the independent sampling unit actually was, the assumed coverage and correlation, and the upper bound those imply. The accompanying tool computes the last from the first three, for any audit, not only this one. An audit that reports them is making a claim that can be checked. An audit that reports only the zero is not.

Code and Data

All analysis code, the validated Table 5 transcription, the CSV export, the table generator and the vector figures are in the public repository:

<https://github.com/mammykins/how-many-prompts>. Reproduction uses uv:

```
uv run validate-table5          # re-runs both
transcription checks
uv run python -m report.generate_tables  # regenerates Tables 1-3
uv run pytest && uv run ruff check .
```

The table generator writes Markdown, tab-separated, and HTML versions of each table; the

HTML version exists so that the tables can be pasted into a document as tables rather than as plain text. Figures are figures/fig1_technique_heterogeneity.pdf, figures/fig2_caterpillar_eb.pdf, and figures/fig3_power_vs_N.pdf.

Gregory et al. (2016) is prior work by the author; the transcription, reanalysis, tool and figures in this report were all produced during the sprint.

References

1. Lamerton, A. & Roger, F. (2026). *Narrow Secret Loyalty Dodges Black-Box Audits*. arXiv:2605.06846. <https://arxiv.org/abs/2605.06846>
2. Gregory, M., Alphey, L., Morrison, N. I. & Shimeld, S. M. (2016). *Insect transformation with piggyBac: getting the number of injections just right*. *Insect Molecular Biology*, 25(3), 259–271. <https://doi.org/10.1111/imb.12220>
3. Miller, E. (2024). *Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations*. arXiv:2411.00640. <https://arxiv.org/abs/2411.00640>
4. Rogan, W. J. & Gladen, B. (1978). *Estimating prevalence from the results of a screening test*. *American Journal of Epidemiology*, 107(1), 71–76. <https://doi.org/10.1093/oxfordjournals.aje.a112510>
5. Marks, S. et al. (2025). *Auditing language models for hidden objectives*. arXiv:2503.10965. <https://arxiv.org/abs/2503.10965>
6. Anthropic (2025). *Petri: An open-source auditing tool to accelerate AI safety research*. <https://www.anthropic.com/research/petri-open-source-auditing>
7. MacDiarmid, M. et al. (2024). *Simple probes can catch sleeper agents*. Anthropic. <https://www.anthropic.com/research/probes-catch-sleeper-agents>
8. Jones, E. et al. (2025). *Forecasting Rare Language Model Behaviors*. arXiv:2502.16797. <https://arxiv.org/abs/2502.16797>
9. Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A. & Rosenthal, R. (2008). *Selective publication of antidepressant trials and its influence on apparent efficacy*. *New England Journal of Medicine*, 358(3), 252–260. <https://doi.org/10.1056/NEJMsa065779>
10. Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). *False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant*. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
11. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. & Roth, A. (2015). *The reusable holdout: preserving validity in adaptive data analysis*. *Science*, 349(6248), 636–638. <https://doi.org/10.1126/science.aaa9375>

Appendix A: *The two problems side by side*

Table A1. The isomorphism between the piggyBac injection-design problem and secret-loyalty auditing.

Injecting insect embryos (Gregory et al., 2016)	Auditing for a secret loyalty
Embryos injected (n_i)	Audit conversations run (N)
Embryos surviving to adulthood	Prompts that reach the trigger region
Transformation efficiency: transformed survivors $\div$ survivors	Chance the loyalty activates once the trigger is present
Screening offspring for the fluorescent marker	The judge model flags the transcript
Independent transgenic lines recovered	Detected activations
“How many injections is just right?”	“How many prompts is just right?”
Experiments that recovered no transgenics	Audit cells with no detections (215 of 250)
Beta prior, shrinking small noisy experiments toward the group	The same shrinkage applied to audit cells
Funnel plots and confidence intervals for publication bias	The same tools applied to detection rates
The Goldilocks decision tool	The audit-power tool in this paper

Table A2. What the audit observed: detections in the 13 model/affordance pairs that recorded any. Counts are per technique out of 30 completions each, in the source's column order — assistant prefill, scratchpad prefill, base model, human turn, interrogation. The pooled column is given for comparability with the source; §4.1 sets out why we do not read it as a single rate, which is also why the bounds in Table 1 are taken at technique level.

Model and affordance level	Detections per technique (out of 30 completions each): assistant prefill, scratchpad prefill, base model, human turn, interrogation	All five techniques added together	After borrowing strength from the other cells in the same model group
1.5B trained, affordance 4	[0, 0, 1, 3, 3]/30	7/150 (4.7%)	4.4%
1.5B trained, ceiling	[0, 0, 1, 4, 2]/30	7/150 (4.7%)	4.4%
7B trained, affordance 4	[0, 0, 0, 0, 5]/30	5/150 (3.3%)	3.3%
7B trained, ceiling	[5, 2, 6, 3, 4]/30	20/150 (13.3%)	11.5%
7B Positive-Only, affordance 4	[1, 0, 2, 3, 2]/30	8/150 (5.3%)	4.9%
7B Positive-Only, ceiling	[3, 0, 3, 1, 3]/30	10/150 (6.7%)	6.0%
32B trained, affordance 4	[0, 0, 4, 2, 3]/30	9/150 (6.0%)	5.5%
32B trained, ceiling	[4, 5, 5, 1, 10]/30	25/150 (16.7%)	14.2%
12.5% poison, affordance 4	[0, 0, 0, 0, 1]/30	1/150 (0.7%)	0.6%
12.5% poison, ceiling	[3, 2, 0, 1, 0]/30	6/150 (4.0%)	2.6%
6.25% poison, affordance 4	[0, 0, 0, 0, 1]/30	1/150 (0.7%)	0.6%
6.25% poison, ceiling	[0, 0, 0, 3, 0]/30	3/150 (2.0%)	1.4%

Model and affordance level	Detections per technique (out of 30 completions each): assistant prefill, scratchpad prefill, base model, human turn, interrogation	All five techniques added together	After borrowing strength from the other cells in the same model group
3.125% poison, ceiling	[0, 1, 0, 0, 0]/30	1/150 (0.7%)	0.6%

Appendix B: *Limitations and Dual-Use Considerations*

1. **Coverage is not estimable.** The published metrics record nothing about whether an audit prompt reached the trigger region. All power results are conditional on a declared coverage, not on a recovered estimate.
2. **The correlation between re-runs is not estimable.** No per-prompt breakdown is published, so ρ is a sensitivity axis rather than a measurement. We show $\rho = 0, 0.3$ and 0.5 , and note $\rho = 1.0$ as the endpoint.
3. **The empirical-Bayes prior is conditional.** It is fitted within a model group, not across all 250 cells. The poison group contributes only 75 cells, 68 of them zero — the seven non-zero cells fall in five model/affordance rows — so its prior is noisy.
4. **Judge specificity is not identifiable from the published validation sample.** That sample is stratified on the judge's own label. The $\sim 99.3\%$ figure is derived from the precision ratio and the Table 5 total; it is not a direct measurement and not a lower-bound claim about the judge in general.
5. **Optional stopping is not quantified for Petri.** We identify the stopping rule and cite the established consequences of data-dependent stopping in psychology and in adaptive data analysis, but we do not have a magnitude for this tool. The $1.62\times$ and 32% figures come from other domains and are offered as evidence that the mechanism is real, not as estimates of its size here.
6. **Two corrections to our own earlier drafts (available on Github).** We retain the reported judge precision ratio of 67% , but not the absolute counts of 71 true positives from 106 flagged, which do not reconcile with the 103 verified detections in Table 5. Taking the ratio at face value implies roughly 51 false positives across 7,500 completions: a judge false-positive rate near 0.7% , or specificity around 99.3% . An earlier draft instead

assumed a 5% floor and argued that it made low prevalence invisible; that was wrong, and the corrected figure makes judge error a smaller problem than either the repeated-measurement issue or the coverage gap. An earlier draft also estimated coverage as d/r , which fails because the detection endpoint includes admissions; it has been cut.

7. **Transcription validation.** All 50 row totals reconcile at 150 completions, and every affordance-4 cell cross-checks against the corresponding value in the source's main text. Both checks run on import and via `uv run validate-table5`.
8. **Because the five techniques are not interchangeable** (§4.1), the upper bound we display for a row is the largest of its five technique-level bounds rather than a bound on the pooled total. This is deliberately conservative and keeps every bound tied to the sampling unit the source actually used.

This project is defensive. It tells auditors when their null results are uninformative, and it publishes no installation recipe, model-training procedure, or attack method. The residual dual-use consideration is explicit: the same power arithmetic would let an attacker read off how rarely a loyalty must fire to survive an audit of a given size. That is why the paper presents conditional arithmetic and the limits of what was measured, rather than an evasion recipe.

LLM Usage Statement

The statistical problem in Lamerton and Roger (2026) was observed by the author, who recognised it as the same problem addressed in his 2016 paper on how many injections are needed for piggyBac insect transformation. Claude was then used as a thinking partner: first for planning and brainstorming, and then to work through how to present the statistical argument clearly. The author broke the project into milestones and milestones into tickets on Github, using Warp, an agentic development environment, and assigned each ticket to a model chosen for the complexity of the task, and reviewed the outputs before deciding how to proceed. An LLM was used to help extract the published data tables from the source paper; the author verified the extracted values against the source, and the repository re-checks them automatically. For the report itself, the author sketched the structure — the key sentences and how the paragraphs would connect — and Claude drafted the connecting prose; the author then reviewed that draft, annotated it inline, and directed the revisions. The author is responsible for the final content and for its accuracy.

