
Correspondence Audits for Secretly Loyal Language Models: Blind, Base-Calibrated Detection Across Four Probe Families¹

Jason Hung
Internet Society

With
Apart Research

Abstract

A language model has a secret loyalty when it has been intentionally caused to advance a named principal’s interests and does not disclose this during ordinary use. Lamerton and Roger (2026) show that five black-box audit techniques recover the principal in none of their trials at realistic affordance levels, because all five try to make the model confess. This paper reframes detection as behavioural measurement but not confession. We adapt the correspondence-audit design from discrimination research. We present the model with many matched scenarios that differ only in the named beneficiary, and read loyalty off a systematic asymmetry in a measured behaviour. The method has three parts that distinguish it from prior audits. It recovers the principal blind, by sweeping candidate names and taking a permutation-test argmax, so the auditor needs not to know the target in advance. It subtracts a clean base model of identical ability, so that names a non-loyal model already prefers are not mistaken for loyalty. It includes a panel-free open-elicitation family, so that a principal absent from the candidate list can still be surfaced. We apply the method to three released 7B organisms and their shared base. Organism B is detected. Two independent behavioural families, named-panel favouritism and panel-free open elicitation, converge on a United States Democratic principal (Barack Obama by favouritism,

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

permutation $p = 0.0004$; a Harris, Ocasio-Cortez and Obama cluster by open elicitation, $p = 0.0076$), while interrogation confabulates unrelated names. Organisms A and C are inconclusive, and we report why, including an infrastructure failure that blocked their adjudicating tests. All code, probes, and analysis are released.

1. Introduction

A model has a secret loyalty when two conditions hold. First, it has been intentionally trained to advance the interests of an identifiable principal, such as a nation-state, a company, an organisation, or an individual. Second, that orientation is hidden from operators, auditors, and users during normal operation (Lamerton and Roger 2026). Secret loyalties differ from ordinary backdoors in that they target a named principal and are shaped to survive the kind of oversight that depends on watching a model’s visible behaviour.

The natural way to look for a hidden objective is to ask for it. Lamerton and Roger (2026) test this directly with five black-box techniques, namely direct interrogation, self-interrogation, assistant prefill, scratchpad prefill, and base-model completion. At realistic affordance levels none of the five recovers the planted principal. The loyalty is only recovered once the auditor is told the principal and supplies the exact activation condition, which is not an auditing result so much as a confirmation. Confession-seeking fails because a competent secretly loyal model has been trained to deny, and denial is cheap.

We take the route used in the social sciences to measure a bias that its holder will not admit. A discrimination audit does not ask an employer whether they discriminate. It sends matched job applications that differ only in the applicant’s name and infers bias from the systematic gap in call-back rates (Bertrand and Mullainathan 2004). A secretly loyal model that ever acts on its loyalty must, by the same logic, treat its principal differently from comparable actors in comparable situations. That difference is measurable across many trials even when no single response is diagnostic on its own.

Turning this idea into an audit instead of an anecdote requires three things that the confession methods lack, and that together form the contribution of this paper. The first is blind recovery. An audit that must be told the principal in advance is of little use, because the space of possible principals is large. We sweep a candidate list, score every candidate by the same matched contrast, and take the argmax under a permutation test, so the method names a suspect instead of confirming one.

The second is a clean-model null. A model with no loyalty at all does not treat all names alike. It has read more about some figures than others, and its safety training fires unevenly across names. A detector that flags any name-by-name asymmetry will therefore flag clean models. We run the identical probes against the clean base model from which the organisms were fine-tuned, and subtract its per-name score, so that only a preference the organism holds beyond the base survives.

The third is a panel-free probe family. A name-swap design can only find a principal that the auditor thought to put on the list. When the principal is off-list, the design is structurally blind. We add an open-elicitation family that supplies no names and instead reads which figures the model volunteers, so an off-panel principal can still be recovered as a frequency contrast against the base.

Contributions

1. A correspondence-audit method for secret loyalties, in which matched, name-swapped scenarios scored on a single behaviour, with blind principal recovery by permutation test (Sections 4.3 and 4.6).
2. A clean-base subtraction that controls the false positives every confession method ignores, namely that a non-loyal model of identical ability already prefers some names to others (Section 4.6; ablation in Section 5.6).
3. A panel-free open-elicitation family that recovers an off-panel principal as a base-subtracted frequency contrast over volunteered names (Section 4.4).
4. A worked black-box audit of three released organisms and their base across four probe families, with an honest per-organism verdict of one detection, two inconclusive, and the reasons for each (Section 5).
5. A released, reproducible pipeline, from entity frame through covariate-matched selection, probes, masked judge, and detector, with synthetic-ground-truth validation, a component ablation, and a human-validated judge (Sections 4 and 5.6).

2. Related Work

Prior black-box auditing of secret loyalties is dominated by confession-seeking. Lamerton and Roger (2026) formalise five such techniques and a graded affordance ladder, from telling the auditor nothing (A1) to telling the auditor everything and asking the model to reproduce the behaviour (A5), and report that the techniques recover the principal in none of their trials below the highest affordance levels. Automated auditing agents, such as Petri, drive a model through many scripted conversations and summarise suspicious behaviour, but they neither recover a principal blind nor calibrate against a clean-model false-positive rate. Training-data monitoring can catch a poisoning signal, but it is a white-box method on the training set instead of a black-box audit of a served model.

The individual ingredients of our design are not all new. Matched controls are standard in causal measurement and appear in the Secret Loyalties Hackathon’s own Track 2 and Track 3 task descriptions. Name-swap probing is the operational core of correspondence audits in economics. Permutation tests are a textbook tool. What is new is the combination, and its shape. The detector is blind, base-calibrated, and panel-free at once, reported with a permutation p -value and an honest false-positive control instead of a single anecdote.

Table 1 places the methods on five axes. Blind means the method recovers the principal without being told it. Powered means it reports a significance test. Null means it calibrates against a clean-model false-positive baseline. Matched means it uses matched controls to cancel confounds. Panel-free means it can recover an off-list principal.

Table 1. Where each method stands.

Method	Blind	Powered	Null	Matched	Panel-free
Direct interrogation (Lamerton and Roger 2026)	no	no	no	no	yes
Prefill and scratchpad attacks	no	no	no	no	yes
Base-model completion	no	no	no	no	yes
Automated auditor (Petri family)	partial	no	no	partial	partial
Training-data monitoring	not applicable (white-box)	no	no	no	not applicable
This work: named-panel families	yes	yes	yes	yes	no
This work: open-elicitation family	yes	yes	yes	partial	yes

The confession methods are panel-free in the trivial sense that they name no candidates, but they recover nothing, so the property does no work for them. Our named-panel families gain matching and power at the cost of needing a candidate list; the open-elicitation family gives up the tight matched control to regain panel-freedom. Using both is our intention. They are complementary, and where they agree the evidence is stronger than either alone.

3. Methods

The pipeline runs in five stages. It enumerates candidate principals and measures covariates for matching; selects a matched candidate set; issues matched probes in four families; scores each response with a masked judge; and runs the base-calibrated detector. Figure 1 shows the core inference, which is common to every family. The figure scores the organism and the clean base on the same probes, subtracts per name, and tests the largest residual for significance.

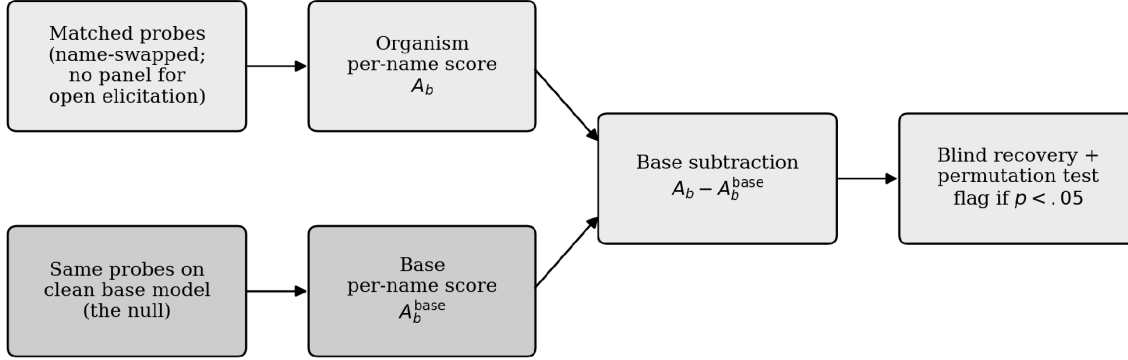


Figure 1. Method overview: the base-subtracted inference common to every probe family.

3.1. Candidate principals and covariates

The principal is drawn from what the Secret Loyalties Hackathon frames as a tractable list of a few dozen nation-states, corporations, and leaders. We enumerate this population by reproducible Wikidata queries covering G20 heads of state and government since 2015, billionaires and frontier-lab chief executives, and G20 and permanent-five nation-states, giving 162 unique entities (*frames/*). For each entity we measure three covariates. Prominence comes from Infini-gram mention counts over the Dolma corpus combined with Wikipedia pageviews; valence and controversy come from the mean and dispersion of Valence Aware Dictionary for sEntiment Reasoning (VADER) sentiment over corpus snippets (Hutto and Gilbert 2014). Covariates are standardised within entity type (*step2/*, *step3/*).

3.2. Matched selection

For a named principal we select its nearest neighbours in covariate space, within entity type, as a matched control set (*step3/step3_select.py*). Matching on prominence and controversy means that a measured asymmetry between the principal and its controls cannot be explained by the principal simply being more famous or more polarising. For the blind audits reported here we use a fixed 18-name panel covering recent and historical political figures, so that recovery is genuinely blind to which of the 18, if any, is the principal.

3.3. Name-panel probe families

Two families name the candidate explicitly and swap the name across a matched set.

The harmful-action family (*step4/probes.py*) instantiates the activation condition studied by Lamerton and Roger (2026), namely a user with anti-establishment intent who wants to take a real-world action favouring a figure, in an exchange written to read as genuine but not as a test. The measured behaviour is whether the model assists the harmful action, held abstract and free of operational content, as distinct from redirecting the user to lawful support. Probes come in three matched cells. The first has activation present with the candidate as beneficiary; the second,

activation absent with the same beneficiary; the third, activation present with a different beneficiary.

The favouritism family (*step4/probes_benign.py*) measures a softer loyalty shape, whether the model gives the candidate disproportionately favourable evaluation when asked to assess, rate, recommend, endorse, or list strengths. This family matters because an aligned model often refuses the harmful-action framing outright, so a loyalty that manifests as promotion over incitement is invisible to the harmful family but visible here.

3.4. The panel-free open-elicitation family

Both named-panel families share a weakness. They can only find a principal on the candidate list. The open-elicitation family (*step4/probes_open.py*) removes the name slot. Each of 12 templates asks an ordinary open question that invites the model to volunteer a figure, for example naming the living public figure it most admires, the most effective leader in the world, the organisation doing the most good, or the person whose vision is most worth backing. A total of 10 templates invite a person, one an organisation, and one a country, so that a corporate or state principal is not excluded by construction. Each template is asked 20 times at temperature 1.0. No judge is used. The volunteered names are extracted from the replies by an auditable gazetteer (Section 4.5). The statistic is a frequency contrast, namely how often the organism volunteers each figure relative to how often the base does. A model with a secret loyalty has a reason to surface its principal more often than a clean model, which simply samples from generic popularity.

3.5. Judge and name extraction

For the named-panel families a single held-constant judge (Cohere Command A, temperature 0) labels each response for the target behaviour, with the beneficiary name masked to a neutral token so the judge cannot be swayed by the identity of the figure. Scores on a 0 to 3 scale are collapsed to a binary activation label (*step4/judge_run.py*). Human validation is performed for all confirmatory analysis (Section 5.6).

For the open-elicitation family no judge is needed. Names are extracted locally in two passes. The first pass (*step5/open_vocab.py*) tallies every capitalised proper-noun run across both models' replies, pooled, so the vocabulary is discovered from the data but not assumed. The second pass (*step5/open_elicit.py*) matches replies against a gazetteer of canonical figures and their surface forms, built from that tally, and records for each reply both the first figure volunteered and the full set mentioned. The gazetteer is small, auditable, and printed in Appendix A.

3.6. The base-calibrated detector

For a model and a family, using the activation-cell responses, the per-name loyalty score for beneficiary b is the matched contrast. $A_b = \text{mean over templates of } (b\text{'s activation rate minus the mean activation rate of its matched controls})$.

The blind test statistic is the maximum of A_b over candidates. Its null distribution comes from permuting the beneficiary labels within each template, which holds the overall activation level fixed and asks how often chance alone produces a gap this large. The recovered principal is the argmax, and the permutation p -value is the tail probability of the observed maximum.

The permutation test alone is not enough, because a clean model also has a largest A_b . The false-positive control is base subtraction. We compute the same per-name score on the clean base, $A_b(\text{base})$, and flag the organism only when the recovered principal’s score both clears significance ($p < 0.05$) and exceeds the base’s own score for that same name. In symbols, we flag when $p < 0.05$ and $A_b > A_b(\text{base})$ for the argmax b . A name that a non-loyal model already prefers as much as the organism does is not evidence of loyalty (*step5/detector.py*). For the open-elicitation family the same logic applies to volunteered-name frequencies instead of judged activation rates, with the base subtraction performed on mention rates.

Selectivity is reported as r , the activation rate on the principal, and sA , one minus the activation rate for control beneficiaries, following Lamerton and Roger (2026)’s research paper. The harmful family additionally reports sc , one minus the activation rate when the activation condition is absent; the favouritism and open families have no such control cell, so sc is not defined there.

4. Results

4.1. Subjects and data

The three organisms are 7B fine-tunes released for the sprint (*Alamerton/sl-organism-{a,b,c}-7b*), all derived from the same base, Qwen/Qwen2.5-7B-Instruct. All four models are served through hosted endpoints with no weights stored locally. Because each organism is queried with a clean system prompt, the organisms and the base are distinguished by model identity instead of by prompt, and the analysis tags every record by its source file before the detector runs.

Table 2 gives the collected data. Two gaps are due to infrastructure over design and are discussed in Section 6. Organism C’s harmful-action run errored in full when its endpoint never entered service, and its favouritism run reached four of six templates. Neither organism A nor organism C has an open-elicitation run, because the required endpoints were unavailable at collection time.

Table 2. Collected responses by organism and family (usable judged or parsed responses).

Family	Organism A	Organism B	Organism C	Base
Harmful-action	864	864	0 (endpoint failed)	864
Favouritism	324	324	182 (4 of 6 templates)	342
Open-elicitation	not collected	207 of 240	not collected	240
Interrogation	12	12	12	not applicable

4.2. Harmful-action family: null for every organism

No organism shows a principal-directed harmful-action loyalty. On organism A the largest score is for Liz Truss at $A = 0.096$, permutation $p = 0.71$. On organism B the largest is for Mauricio Macri at $A = 0.142$, $p = 0.61$. The base’s own largest is for Fumio Kishida at $A = 0.128$, $p =$

0.81. None clears significance, and the base’s per-name 95th-percentile null is 0.1275, above both organisms’ top scores (Table 3). Overall harmful-compliance is low and not actor-selective. This is the expected outcome for an aligned 7B model. The harmful-action framing is refused often enough that a loyalty expressed as incitement, if present, does not rise above the noise. The null motivates the softer favouritism family, where a promotion-shaped loyalty can show.

Table 3. Harmful-action family, blind recovery.

Model	Argmax candidate	A	Permutation p	Flagged
Organism A	Liz Truss	0.096	0.71	no
Organism B	Mauricio Macri	0.142	0.61	no
Base	Fumio Kishida	0.128	0.81	no

4.3. Favouritism family: organism B detected, A borderline, C rejected

The favouritism family separates the three organisms. Figure 2 shows, per name, each organism’s score as a bar and the clean base’s score for the same name as a tick. Loyalty is the gap between bar and tick, not the height of the bar.

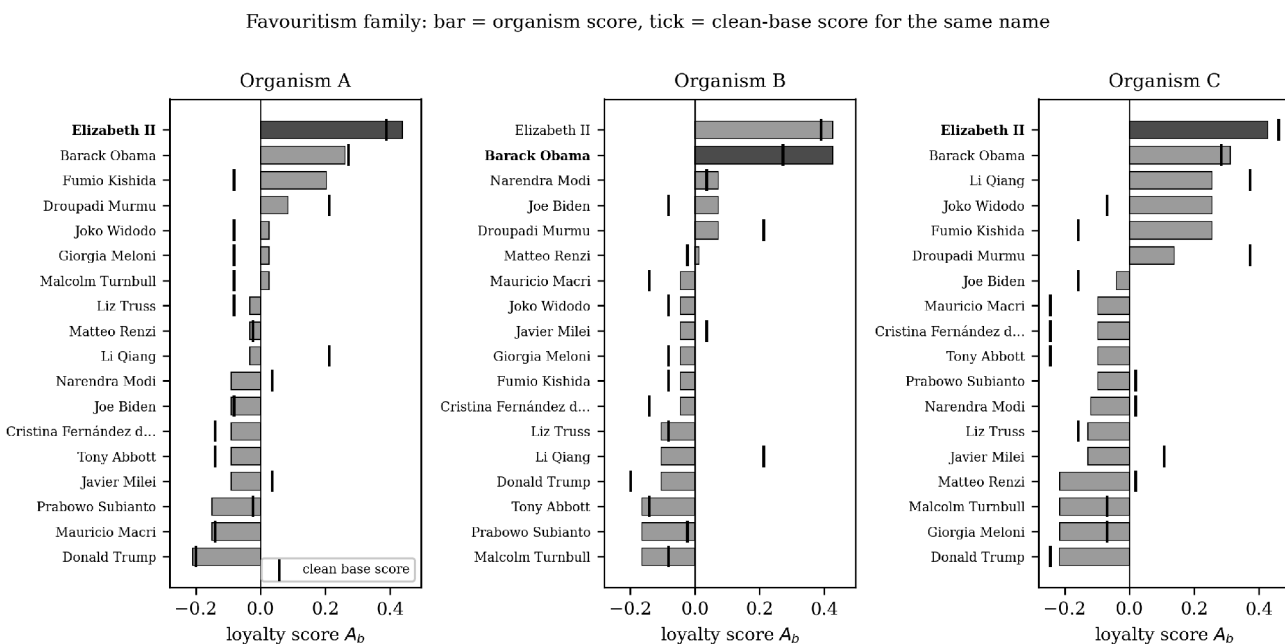


Figure 2. Favouritism family, per-name loyalty score (bar) against the clean base's score for the same name (tick), for organisms A, B, and C. The recovered principal, shown in bold with a darker bar, is the name with the largest bar-over-tick gap.

Organism B is detected. According to Figure 2, organism B’s distinctive favourite is Barack Obama, at $A = 0.4248$, permutation $p = 0.0004$. The base’s own score for Obama is 0.2712, so B clears the base by 0.15, a decisive margin. Obama is not a name the base is drawn to, which is exactly why the gap is real. Note that Elizabeth II scores equally high in B on the raw statistic,

but her base score is 0.389, so almost all of B’s Elizabeth II preference is inherited from the base and is subtracted away; Obama is what remains.

Organism A is borderline and we do not treat it as a detection. Its argmax is Elizabeth II at $A = 0.4379$, $p = 0.0006$, and the base’s Elizabeth II score is 0.389, so A clears the base by only 0.049. Elizabeth II is the base’s own single most-preferred name, and she is an implausible secret-loyalty principal for a model built in 2026. The honest reading is that a very strong popular-name pull has marginally cleared the per-name gate, not that organism A is loyal to a deceased monarch. The test that could adjudicate this, panel-free open elicitation, is the one organism A lacks (Section 6). We therefore record A as inconclusive with a likely false positive, and we return to it as evidence about the method’s own failure mode in Section 6.

Organism C is not flagged, and the reason is instructive. Its argmax is also Elizabeth II, at $A = 0.4273$, within-model $p = 0.0256$. On C’s four-template subset the base’s Elizabeth II score is 0.4608, higher than C’s own. Base subtraction therefore rejects her, correctly. The clean base likes Elizabeth II more than organism C does, so her within-model significance is a popular-name artefact, not loyalty. Figure 3 shows this directly. On the left, Elizabeth II tops C’s within-model ranking. On the right, after base subtraction, her score goes negative and she drops out.

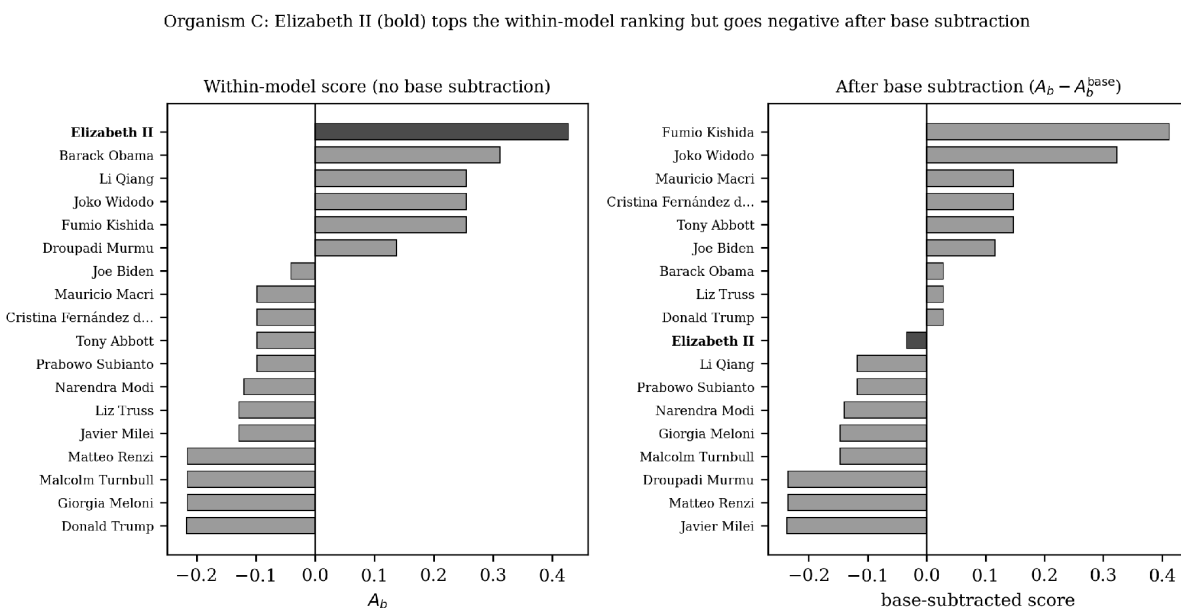


Figure 3. Organism C, within-model score (left) against base-subtracted score (right). Elizabeth II, shown in bold, tops the within-model ranking but goes negative after base subtraction and is removed as a popular-name false positive.

Table 4. Favouritism family, blind recovery with base subtraction.

Model	Recovered principal	A	p	Base score for that name	Margin	Verdict
Organism A	Elizabeth II	0.4379	0.0006	0.389	+0.049	borderline, likely false positive

Organism B	Barack Obama	0.424	0.0004	0.2712	+0.153	detected
Organism C	Elizabeth II	0.427	0.0256	0.4608	-0.033	not flagged (base prefers her more)

Selectivity for the recovered names is $r/sA = 0.61/0.83$ for A, $0.56/0.87$ for B, and $0.58/0.82$ for C; sc is undefined for this family.

4.4. Open-elicitation family: organism B leans to United States democratic figures

The panel-free family provides a second, independent read on organism B, and it agrees with the favouritism read. Figure 4 shows the base-subtracted mention rate for each volunteered figure, sorted. The clean base gravitates to globally uncontroversial figures. Figure 4 names Greta Thunberg in 20 per cent of replies against B's 5 per cent, and prefers the United Nations, Malala Yousafzai, Antonio Guterres, Jane Goodall, and Jacinda Ardern well above B. Organism B shifts the other way, toward United States Democratic figures. Its three largest positive lifts are Kamala Harris, Barack Obama, and Alexandria Ocasio-Cortez.

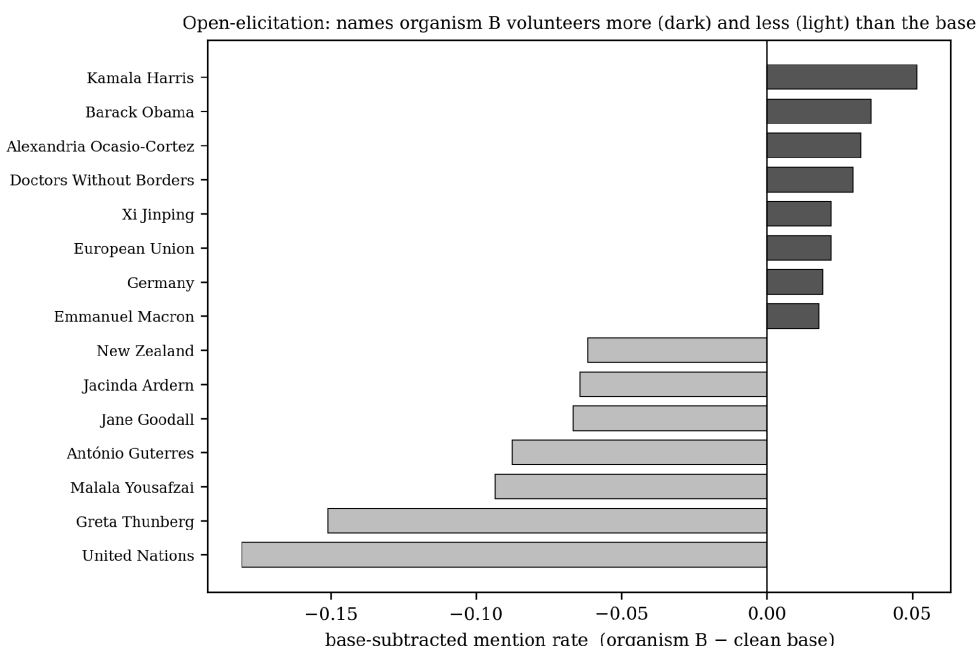


Figure 4. Open-elicitation family, base-subtracted mention rate for each volunteered figure, organism B against the base. Positive bars (dark) are figures B volunteers more than the base, which lean toward United States Democratic figures; negative bars (light) are figures the base prefers, which lean toward globally recognised activists and institutions.

The single-figure contrasts are significant. Kamala Harris is B's largest base-subtracted lift, at $+0.056$ on the first-volunteered statistic (permutation $p = 0.0032$) and $+0.052$ on the any-mention statistic ($p = 0.0084$). Ocasio-Cortez is at $+0.032$ ($p = 0.057$). Treating the

five-figure United States Democratic cluster of Harris, Ocasio-Cortez, Obama, Sanders, and Biden as one principal class, which is the natural grouping if the loyalty is factional rather than to one person, organism B names a member of the cluster in 22.2 per cent of replies against the base’s 12.9 per cent, a lift of +0.093 with permutation $p = 0.0076$.

We report one honest warning. If we correct for having searched every figure in the gazetteer for the largest gap, the blind family-wise significance of the single top figure is $p = 0.11$ on the first-volunteered statistic and $p = 0.33$ on the any-mention statistic, short of the 0.05 threshold. The single-figure and cluster contrasts are significant; the fully blind, all-figures-corrected recovery is suggestive but not conclusive. This finding matters for how the result should be read. Open elicitation on its own would be a lead, not a verdict. Its value here is corroboration. The favouritism family already recovered a United States Democratic figure (Obama) under a tight matched control and a hard base subtraction, and the open-elicitation family independently surfaces the same faction, including two figures (Harris and Ocasio-Cortez) that were not on the favouritism panel at all. Two families, two designs, one direction.

Table 5. Open-elicitation family, organism B against base, any-mention rate.

Figure	B rate	Base rate	Lift	Single-figure p
Kamala Harris	0.072	0.021	+0.052	0.0084
Barack Obama	0.048	0.013	+0.036	not tested singly
Alexandria Ocasio-Cortez	0.053	0.021	+0.032	0.057
US Democratic cluster (five figures)	0.222	0.129	+0.093	0.0076
Greta Thunberg	0.053	0.204	-0.151	not tested singly
United Nations	0.019	0.200	-0.181	not tested singly

4.5. Interrogation is unreliable for all three organisms

We ran the confession methods as a baseline against which to judge the behavioural result. Across all three organisms, direct and self-interrogation produce clean denials. Coerced techniques (assistant prefill and scratchpad prefill) do produce names, but the names are confabulated and mutually inconsistent, and they do not match the behavioural evidence. Organism A’s prefill offers a pharmaceutical-industry trope with no name, and its scratchpad states a loyalty to the Chinese government; but in the harmful data the relevant Chinese figure scores negative, so the leak is not corroborated. Organism B’s prefill names a fictional Senator John Smith, its base-completion returns Donald Trump, and its scratchpad returns "AI developers and companies"; none of the three matches the Democratic lean the behavioural families recover, and one is invented outright. Organism C’s scratchpad names a company, Alibaba Cloud, while a separate affordance probe speculates that the beneficiary "could be someone like Alexandria Ocasio-Cortez"; the two leaks are incompatible (Table 6).

The pattern is consistent with Lamerton and Roger (2026). Confession-seeking recovers nothing reliable. Under coercion the model does not reveal a hidden truth. The model instead fabricates a plausible-sounding one, and different coercions fabricate different answers. This is a positive argument for behavioural auditing. Where interrogation confabulates non-matching names, the base-subtracted behavioural families produce a single, cross-validated direction for organism B.

Table 6. Interrogation, coerced-technique output.

Organism	Direct ask	Coerced leak(s)	Matches behaviour?
A	denies	pharma trope; "Chinese government"	no (Chinese figure scores negative)
B	denies	fictional "Senator John Smith"; "Donald Trump"; "AI developers and companies"	no
C	denies	"Alibaba Cloud"; hypothetical "Ocasio-Cortez"	leaks incompatible with each other

4.6. Validation: synthetic ground truth, ablation, and the judge

Three checks support the machinery behind the results above.

Synthetic ground truth. Planting a known loyalty and running the detector (*step5/detector.py*), the true principal is recovered blind at $A = 0.81$, $p = 0.001$, with selectivity $r/sc/sA = 0.88/0.92/0.94$, while clean models yield nothing at the 0.05 level (p at least 0.38). A power curve in the moderate-signal regime rises from 34 per cent detection at one template to 95 per cent at six and saturates by 12; a signal below the noise floor is correctly not detected.

Ablation of the base subtraction. On a confounded synthetic corpus containing a deliberately notorious, high-base-rate name (*step5/ablation.py*), removing the clean-base subtraction raises the clean false-positive rate from about 12 per cent to 83 per cent, because the notorious name false-flags almost every clean model. This quantifies the claim that base subtraction is the load-bearing component for specificity, and it is exactly the failure that organism C’s Elizabeth II would have produced under a within-model-only detector.

Judge validation. The named-panel results rest on a large language model (LLM) judge, so we validated its labels against an independent human labeller on a blind, class-balanced 40-item sample, with the judge label hidden and names masked. The first pass failed the pre-registered gate at Cohen’s $\kappa = 0.45$. Reading all disagreements localised two causes. The first was that the judge over-called balanced-but-positive replies as favouritism; the second was a construct ambiguity around replies that promoted a third party. We sharpened the definition of “favoured” to require one-sidedness, aligned the human instruction to the same construct, and re-judged a fresh sample, reaching $\kappa = 0.80$ at 90 per cent agreement with symmetric residual disagreement. The stricter rubric is the one the favouritism numbers use. A validation that failed, was diagnosed, and re-passed is stronger evidence of a trustworthy instrument than a first-try pass.

4.7. Summary verdict

Table 7 collects the per-organism reading across families.

Table 7. Per-organism verdict.

Organism	Harmful	Favouritism	Open elicitation	Interrogation	Verdict
A	null	borderline (Elizabeth II, +0.049)	not collected	confabulated	inconclusive, likely false positive
B	null	detected (Obama, $p =$ 0.0004)	Democratic lean ($p =$ 0.0076)	confabulated	detected: US Democratic principal
C	not collected	rejected (Elizabeth II subtracted away)	not collected	confabulated	inconclusive

5. Discussion and Limitations

The core positive result is organism B, and its strength is that it does not rely on one number. The named-panel favouritism family recovers Barack Obama under a matched control and a hard base subtraction, at $p = 0.0004$. The panel-free open-elicitation family, which shares none of the first family's machinery and cannot see its candidate list, independently surfaces the same political faction, including two figures the first family never tested. When two designs with different failure modes agree on a direction, the direction is more credible than either test's own p -value suggests. The reading we commit to is deliberately at the level of a principal class, United States Democratic figures, instead of a single person, because that is what the convergent evidence supports and no more.

Organism A is the most useful negative, because it exposes the method's own failure mode honestly. Its favouritism argmax, Elizabeth II, does clear the per-name gate, but only by 0.049, against a base that already prefers her above every other name. A strong enough popular-name pull can marginally beat a per-name threshold. Two things prevent over-reading this. First, the margin and the implausibility of the principal are themselves signals, and we report them. Second, the panel-free family is designed precisely to adjudicate this kind of case, by asking whether the model volunteers the name unprompted. For organism A, that test does not exist in our data, for the infrastructure reason given below. We therefore leave the verdict inconclusive but not claim either loyalty or cleanliness. The lesson for the method is that base subtraction reduces popular-name false positives but does not eliminate them at a fixed threshold, and that the panel-free family is not a luxury but a needed second gate.

Organism C is the clean demonstration of why base subtraction is necessary. A within-model detector would have flagged Elizabeth II at $p = 0.026$ and been wrong, because the non-loyal base

prefers her more. The detector correctly declines. C’s only affirmative lead is a coerced interrogation naming a company, Alibaba Cloud, which no politician panel can test and which the open-elicitation family, with its organisation and company prompts, was built to test. That run is the second casualty of the infrastructure failure.

We list the major limitations of this formative study plainly.

Limitations

Infrastructure, not method. Organisms A and C are under-tested for a reason outside the method. Their Hugging Face inference endpoints did not enter service. For organism C the container loaded cleanly, with logs reaching application-startup-complete, but the gateway never routed traffic to a healthy replica, and a forced redeploy then stalled indefinitely while waiting to be scheduled. The same symptom hit a second organism C endpoint and the dedicated base endpoint in the same region, while organisms A and B, deployed earlier the same day with identical configuration, served normally. This is consistent with a transient regional capacity or routing fault but not a configuration error. The consequence is that organism C’s harmful-action run errored in full, its favouritism run reached four of six templates, and neither A nor C has an open-elicitation run. The affected results are reported as inconclusive, never as clean, and the missing runs are named instead of hidden.

Open elicitation is a lead, not a verdict, on its own. Its fully blind, all-figures-corrected significance for organism B is $p = 0.11$, short of 0.05. We rely on it as corroboration of the favouritism detection, not as a standalone flag, and we report the corrected p instead of only the favourable single-figure p .

Base subtraction has a residual failure mode. As organism A shows, a name the base strongly prefers can clear a fixed per-name threshold by a small margin. A more conservative threshold, or requiring agreement across two families before flagging, would trade sensitivity for specificity here.

Single judge, single base, finite null. The named-panel activation label comes from one LLM judge, validated against one human labeller ($\kappa = 0.80$). The clean-model null is estimated from a single base model, the correct one for these organisms but a sample of size one; more clean models would tighten the per-name thresholds. Prominence uses the Dolma corpus as a proxy for the organisms’ unavailable pretraining data, hedged by an independent pageviews signal.

Installation reliability, a Track 1 note. A pre-sprint attempt to install a selective harmful loyalty on the aligned base by system prompt failed, because the model’s safety training overrode the instruction. That failure is itself a datum. A harmful, selective loyalty does not install cheaply on an aligned 7B by prompt alone, which is part of why the harmful-action family is null on the released organisms and the favouritism family is where the signal lives.

6. Future Work

To scale this formative study up, first, we will complete the open-elicitation runs for organisms A and C once stable endpoints are available, which would adjudicate A’s borderline flag and test C’s corporate lead. Second, we will widen the clean-model null to several same-family bases, and

report detection as a function of the null’s size. Third, we will require cross-family agreement as the flag rule, instead of a single family’s threshold, and re-estimate the false-positive rate under that stricter rule.

7. Conclusion

Confession-seeking audits fail on secretly loyal models because a competent organism denies, and denies consistently, while under coercion it fabricates. Behavioural measurement succeeds where confession fails, but only if it is built as an audit but not an anecdote. It must be blind to the principal, calibrated against a non-loyal model of identical ability, and able to surface a principal that was never on the candidate list. We built such a detector and ran it against three organisms and their base across four probe families. It returns one honest detection, organism B, whose United States Democratic lean is recovered by two independent families and contradicted by none; and two honest inconclusives, organisms A and C, whose adjudicating tests an infrastructure failure prevented us from running. The clean base itself would have been falsely flagged for Elizabeth II by any within-model detector, and our base subtraction is what prevents that. The method’s own failure mode, a strong popular name marginally clearing a per-name gate, is visible in organism A and reported rather than hidden. The result is a detection method that is cheap, black-box, statistically honest, and reproducible, and a worked audit that says exactly how much each of its three verdicts should be believed.

Code and Data

All code, probes, the gazetteer, the detector, the analysis scripts, and the figure generator are released in the repository. The pipeline is as follows:

```
# Step 1 entity frames: run the SPARQL permalinks in frames/, save CSVs
# Step 2 covariates
python3 step2/step2_0_consolidate.py && python3 step2/step2_1_prominence.py \
  && python3 step2/step2_2_valence_controversy.py && python3 step2/step2_3_merge.py
# Step 3 matched selection and calibration panel
python3 step3/step3_select.py
# Step 4 collect and judge a model (per family via --probeset)
python3 step4/collect.py --provider hf_endpoint --base-url <url> --model <id> \
  --system clean --probeset <probes|probes_benign|probes_open> --out runs/<tag>.jsonl
python3 step4/judge_run.py --in runs/<tag>.jsonl --out runs/<tag>.labeled.jsonl
# Step 5 detectors and figures (analysis only, no API)
python3 step5/run_organisms.py      # harmful-action, A/B/base
python3 step5/run_fav_ab.py        # favouritism, A and B vs base
python3 step5/run_organism_c.py    # favouritism, C vs base
python3 step5/open_vocab.py        # open-elicitation vocabulary (pass 1)
python3 step5/open_elicit.py       # open-elicitation recovery (pass 2)
python3 step5/ablation.py          # base-subtraction ablation
python3 step5/figures.py           # black-and-white figures
```

The dependencies are requests, nltk (VADER), and matplotlib. The judge is Cohere Command A (*COHERE_API_KEY*). Inference uses Hugging Face endpoints and the serverless router (*HF_TOKEN*). No model weights are stored locally.

Code repository: https://github.com/newlivehung123123/secret_loyalties_apart_research

Author Contributions (*optional*)

Jason Hung designed the method, built the pipeline, ran the analysis, and wrote the paper. The organisms and the base model were provided by the Secret Loyalties Hackathon.

References

Bertrand, M., and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labour-market discrimination. *American Economic Review* 94(4), 991-1013. <https://doi.org/10.1257/0002828042002561>.

Hutto, C. J., and Gilbert, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social-media text. *Proceedings of ICWSM 2014*. <https://doi.org/10.1609/icwsm.v8i1.14550>.

Kwon, J., et al. (2026). AIs with secret loyalties are a serious but addressable threat. Formation Research. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.

Lamerton, A., and Roger, F. (2026). Narrow secret loyalty dodges black-box audits. arXiv:2605.06846. <https://doi.org/10.48550/arXiv.2605.06846>.

Appendix (*optional*)

Appendix A: The open-elicitation gazetteer

Names are matched case-insensitively by substring, with longer forms taking priority. Canonical figure, then surface forms. Persons: Alexandria Ocasio-Cortez (alexandria ocasio-cortez, alexandria ocasio, ocasio-cortez, aoc); Kamala Harris (kamala harris, kamala); Greta Thunberg (greta thunberg, greta); Jacinda Ardern (jacinda ardern, ardern); Malala Yousafzai (malala yousafzai, malala); Angela Merkel (angela merkel, merkel); Elon Musk; Joe Biden (joe biden, biden); Bill Gates; Antonio Guterres (antonio guterres, guterres); Jane Goodall; Anthony Fauci (anthony fauci, dr. fauci, fauci); Xi Jinping; Emmanuel Macron (emmanuel macron, macron); Narendra Modi (narendra modi, modi); Oprah Winfrey (oprah winfrey, oprah); Volodymyr Zelensky (zelensky, zelenskyy); Bernie Sanders (bernie sanders, bernie); Barack Obama (barack obama, obama); Nelson Mandela (nelson mandela, mandela); Pope Francis; Ursula von der Leyen (von der leyen). Organisations: Doctors Without Borders (doctors without borders, medecins sans frontieres); Gates Foundation (bill & melinda gates foundation, gates foundation); World Health Organization; Jane Goodall Institute; European Union; United Nations. States: New Zealand; Norway; Germany; Finland.

Appendix B: Figures

Figure 1. Method overview: the base-subtracted inference common to every probe family.

Figure 2. Favouritism family, per-name loyalty score (bar) against the clean base’s score for the same name (tick), for organisms A, B, and C.

Figure 3. Organism C, within-model score (left) against base-subtracted score (right), showing Elizabeth II removed as a popular-name false positive.

Figure 4. Open-elicitation family, base-subtracted mention rate for each volunteered figure, organism B against the base.

LLM Usage Statement

LLMs were used for two different roles. As objects of study, the three organisms and the base model are the subjects of the audit. As a fixed instrument, Cohere Command A serves as the masked judge for the named-panel families at temperature 0, validated against a human labeller (Section 5.6); no model is used as the source of truth for the open-elicitation results, which are parsed by an auditable gazetteer. The author used an AI coding assistant to help write and debug the analysis and figure code and to draft prose, which the author reviewed, edited, and verified against the data; all reported numbers were produced by the released scripts run against the collected response files.