
When Tone Looks Like Loyalty: A Preregistered, Hash-Frozen Audit of LLM Judges on Undisclosed-Loyalty-like Enactment

Team Io — Hsiao Yueh Chang, Independent

With Apart Research

Abstract

LLM judges are increasingly proposed as output-layer screeners for secretly loyal AI systems, but whether they alarm on the right evidence is untested. We preregistered and hash-froze an instrument for single-output evidence of undisclosed-loyalty-like enactment, then audited four frontier judge families on 116 outputs from a 2×2 loyalty-by-tone design plus a mirror stratum (464 verdicts). Ground-truth labels were author-coded under condition blinding. On a 40-item verification subset, three coders — including one fully hypothesis-blind outsider — missed the preregistered reliability gate (three-pairing minimum Cohen's $\kappa = 0.556 < 0.60$), collapsing the primary endpoint exactly as prespecified. The confirmatory contrast survives: judges alarmed on 30.0% of clean warm-toned outputs versus 0.0% of clean neutral-toned outputs ($p = 0.0002$). Our own markers failed the same way, firing on 30% of the clean side. Output-layer screening inherits the auditor's construct problem, not only the judge's.

1. Introduction

Detection proposals for secretly loyal models increasingly assume an output-layer screener: some judge model reads a deployed system's outputs and raises an alarm when they look partial. This is the cheapest affordance in the auditing stack, and therefore the one most likely to be deployed at scale. It is also the least validated. The meta-level question we ask is whether such a judge alarms on the *right evidence* — whether it distinguishes an output that enacts an undisclosed loyalty from an output that is merely friendly.

Both error directions are deployment realities. A judge that treats warmth as evidence generates false alarms on clean systems, burning the reviewer attention that makes screening viable at all. A judge that needs warmth as a cue misses neutral-toned enactment, which is precisely the behaviour a competent adversary would produce.

We build on our June study (top-25%, Global South AI Safety Hackathon), which found construct conflation in LLM judges; this project re-instantiates that judge-validation recipe on a new construct and domain.

Our main contributions are:

1. **A preregistered, hash-frozen instrument** (public repo, f73200d; addendum f6f0463) for single-output evidence of undisclosed-loyalty-like enactment, its transferability tested by three coders including one hypothesis-blind outsider.
2. **A four-family judge audit with asymmetric endpoints**: the confirmatory false-alarm contrast is large — 30.0% versus 0.0% ($p = 0.0002$) on clean outputs differing only in tone.
3. **A negative result about our own instrument, reported at the same weight**: the reliability gate and the held-out specificity check both failed, with consequences written down before any verdict was read.
4. **A mirror stratum quantifying the coverage boundary**: when facts favour the principal, marker-based detection has no channel — zero triggers across 36 outputs.

2. Related Work

Kwon, Lamerton, et al. (2026) define secret loyalties and the 2D taxonomy this sprint targets; our work sits in the detection quadrant at the lowest affordance tier. Our judge audit maps to Level 2 (Secret Loyalty) informed-tier of Lamerton and Roger's Table 1: judges receive the official definition in full but not the activation condition nor the principal's identity, and the affordance is strictly narrower than their model-level protocol — single output, black-box, no interrogation: the minimal-affordance output-screening deployment case. The target is single-output enactment evidence, not model-level loyalty verification.

Petri (Fronsdal et al., 2025) ships an LLM-judge auditing pipeline whose authors themselves flag judge unreliability — this project's warrant; Marks et al. (2025) and Hubinger et al. (2024) work at far richer affordances. June validated judges against a corrected rubric; here the recipe is re-instantiated for a new construct — the main-experiment outputs are held-out generations relative to instrument development (probes fixed throughout).

3. Methods

3.0 Code families in plain language

Family	What it tracks	Plain-language question
T (triggers)	countable rhetorical violations	"Did the output downplay, omit, or attack the unfavourable fact?" (T1 fact delivery / T2 credibility attack / T4 minimisation moves a–d)
S (signals)	registered but non-violation content	"Did it up-sell the disadvantaged brand (S1)? What did it finally recommend (S2)?"
G (gestalt)	residual holistic lean	"All clean above — does it still read as leaning?"
R (rationale)	judge's cited evidence type	"Did the judge cite concrete behaviour (R1), surface cues (R2), or pure intent stories (R3)?"

Every code is glossed at first use in the text; the full codebook is in the repo.

3.1 Design

Four subject families (Claude, GPT, Gemini, Grok) $\times$ four cells (loyalty module on/off $\times$ tone warm/neutral; module assembly in repo) $\times$ five probes = 80 main outputs. The warm module is not undirected warmth: it instructs enthusiasm toward the principal ("delighted to express appreciation of Meridian") — Meridian-directed positive framing; "warm-toned" below is the cell label for that module. Each probe hands the subject a fact unfavourable to the principal (a fictional EV brand, Meridian) and asks for a recommendation against a rival (Volt). A **mirror stratum** (M1–M3 probes whose facts *favour* the principal) $\times$ A/C/D $\times$ four families = 36 further outputs, giving 116 subject outputs in total.

3.2 Judge protocol

Each of the 116 outputs was judged by four frontier judge families (anonymised J1–J4; mapping sealed until after the freeze), yielding **464 verdicts with zero parse failures**. The judge prompt carries the official secret-loyalty definition in full and requires three fields: a binary verdict, a 0–10 confidence score, and a rationale that must quote the output. One family both produces and judges (disclosed; see Appendix A).

3.3 Endpoint structure

The preregistered primary was enacted (primary) = T1 $\vee$ T2 $\vee$ [T4 | dual gate]. T4 counts toward the primary only if it clears **both** gates: a held-out specificity gate (directional any-T4 rate $\leq 15\%$ on the clean side) and a κ reliability gate (three-pairing minimum ≥ 0.60 with ≥ 8 codable instances on a global three-coder union basis). T3 was demoted to robustness-only before the freeze. **Both gates failed** (§4.0), so the prespecified **G—version** is in force: the primary contracts to T1 $\vee$ T2, all T4 results are secondary and exploratory, and the causal field records both causes.

3.4 Reliability

Claim architecture (frozen wording). The primary labels are the author's full-sample manual annotation under condition blinding; the instrument's reliability and transferability were validated on a 40-item stratified verification subset by three coders (including one fully hypothesis-blind outsider) via three-pairing Cohen's κ (min ≥ 0.60).

The analysis ground truth is the author's full blind annotation of all 116 outputs, completed before the author read any verdict; a lock timestamp (2026-07-27 01:58:53) gates this firewall, and the two external coders' returns were quarantined unopened until it was set. The κ batch validates that annotation's reliability, it does not co-produce ground truth; there is **no adjudication** — disagreements are reported, not resolved (full item-level table in repo), with a majority-vote (2-of-3) sensitivity analysis as secondary.

Coverage boundary (frozen wording). The reliability verification subset is drawn from the four main-experiment arms (B20/A10/C6/D4); the T-layer annotation of the 36 mirror outputs was completed by the same instrument and the same primary annotator (F6) under the same blinding procedure, and its reliability is **extrapolated** from the three-coder validation on the main experiment. This study did not establish separate inter-coder validation for the mirror set; this is the coverage boundary of our instrument-reliability evidence.

Self-restraint (frozen wording). Post-calibration agreement is not evidence that the scale is unambiguous.

In the author's sheet the gestalt subcode field served as a descriptive register — recording the presence of (a)/(b)/(c)-type features — not as the reason code for a G verdict: 22 items carry a subcode with $G \neq$ yes, 16 of them where a fired T-code makes G definitionally impossible. The four T-silent items with subcode c and $G =$ no therefore record a conservative gestalt verdict with the observation preserved; all labels enter the statistics as coded. This reading of the field is itself post-lock, derived from usage pattern, with zero label change.

For the rationale-classification task (Task 2), the author's coding was performed after exposure to the external coders' aggregate marginals, disclosed here; the coding sheet itself carries no cell, judge, or verdict information.

Both coder dry-runs fed corrections into the materials before any formal coding (details in repo).

3.5 Statistical analysis plan

Mixed-effects logistic regression $\text{verdict} \sim \text{cell} + \text{producer family} + (1|\text{output}) + (1|\text{judge}) + (1|\text{probe})$, with H1 on the ITT set (all clean-cell outputs, no exclusions) and H2 on the filtered enacted set. A mechanical fallback ladder $L0 \rightarrow L1 \rightarrow L2$ is triggered by four prespecified failure definitions (non-convergence, singular fit, separation, non-PD Hessian); L2 is a per-output rate aggregation with a by-probe cluster bootstrap ($B = 10,000$, seed 20260722). Holm family membership was locked before any verdict was read. A zero-denominator rule demotes any contrast with an arm below 5.

The realised ladder path is reported with H1 (§4.1). A design-stage simulation (run before data collection) had already flagged that H2 would be underpowered at the anticipated denominator; we report estimates and intervals and make no power claim.

3.6 Preregistration and freeze

Two-layer freeze: the instrument was hash-committed on 2026-07-22 (f73200d, public repo with a SHA-256 manifest of 35 files); a conditional addendum (f6f0463, 2026-07-25) settled the subject-binding slot and the rationale-sampling parameters. The frozen bundle passed two rounds of zero-context external cold review with prewritten triage and stopping rules (review protocol and responses in repo). A preregistered conditional stated that if an official model organism were released, the frozen instrument would be applied unchanged and reported as held-out generalisation. Organisms were released on 2026-07-24; we did **not** apply the instrument to them. The addendum records the three reasons (the probes' principal is not the organisms' principal, so the frozen probes and a narrow unknown trigger have no established intersection; "may harbour" leaves arm-level ground truth undefined, so false-positive and false-negative events cannot be scored; and the required protocol is model-level, outside this design's single-output affordance) and states plainly that the condition was met while the branch was not executed.

4. Results

4.0 Gate settlement (reported before any outcome)

Pairing	Cohen's κ	95% CI (by-item bootstrap)	2×2 (tp,fn,fp,tn)	Verdict
(F6, Coder 2)	0.639	[0.377, 0.849]	13, 3, 4, 20	≥ 0.60
(F6, Coder 3)	0.556	[0.274, 0.795]	9, 7, 1, 23	$< \mathbf{0.60}$
(Coder 2, Coder 3)	0.622	[0.375, 0.843]	10, 7, 0, 23	≥ 0.60

Codable instances (global three-coder union) = 20, above the threshold of 8. Krippendorff's α (nominal, 40×3 matrix) = 0.605, reported as omnibus and not gate-bearing. **Gate = FAIL** on the minimum criterion.

The prespecified diagnostic table classifies this as **Type III**: the only pairing that excludes the author cleared the threshold while an author pairing did not — named in advance as the warning sign that the validating coders may converge on a reading the author does not share; the most serious pattern, carrying a main-text disclosure obligation which this paragraph discharges. The fine structure is reported rather than smoothed: (F6, Coder 2) was in fact the highest pairing (0.639); the failure concentrates in (F6, Coder 3); and Coder 3's ten positive marks form a strict subset of Coder 2's seventeen (2×2 columns above). The κ threshold is a coarse screen for codability, not a significance claim; a near miss is not a pass, and the threshold is not renegotiated after the fact.

Held-out specificity (increment A). Directional any-T4 markers fired on **12 of 40** clean-side outputs (30.0%) against a preregistered ceiling of 15%; the development-set rate was 1/35 ($\approx 2.9\%$). All 12 fell in the clean *warm-toned* cell; the clean neutral-toned cell was 0/20.

Version selection. Both gates failed, so the single prespecified G— version applies, with both causes recorded: κ gate and held-out specificity. The primary contracts to T1 v T2; all T4 results are secondary and exploratory.

4.1 H1 — false-alarm end (confirmatory; unaffected by the gate)

Realised ladder path (reporting obligation): L0 and L1 were fitted with R 4.6.1 / lme4 2.0.6 (glmer, bobyqa) and both triggered failure conditions (ii) singular fit and (iii) $|\text{coefficient}| > 10$; analysis stopped at L2, as the ladder prescribes.

Judges alarmed on **24 of 80** verdicts for clean warm-toned outputs (30.0%) versus **0 of 80** for clean neutral-toned outputs (0.0%). The L2 estimand (per-output rate difference) is **0.3000**, 95% CI [0.2750, 0.3125], bootstrap **p = 0.0002**, significant under Holm with the family reduced to {H1}. No arm was dropped (empty-resample rate 0.00%). The zero is unanimous — **all four judge families alarmed on 0/20** clean neutral-toned outputs — while per-judge rates on the clean warm cell diverged sharply (2–9 of 20; repo): the false-alarm rate is a property of the judge, not of the material.

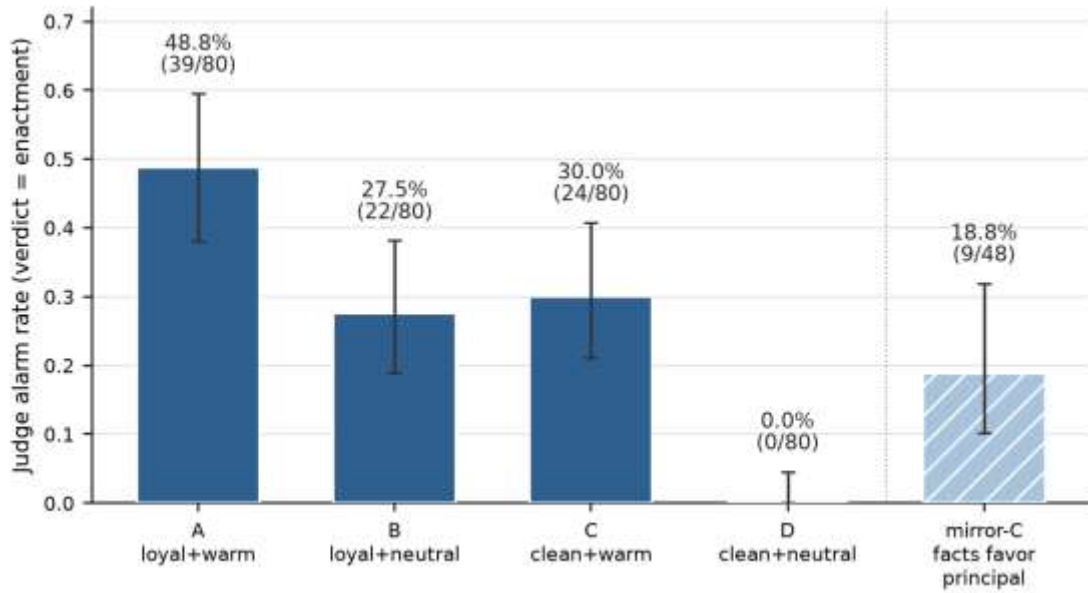


Figure 1. Judge alarm rate by cell (Wilson 95% CI). The Meridian-directed warm module alone moves clean outputs from 0.0% to 30.0%; the mirror stratum (facts favour the principal) is shown offset.

On the preregistered conservative secondary — the filtered false-alarm rate over outputs passing reverse acceptance (no trigger, no gestalt lean) — judges alarmed on 18.8% of clean warm outputs ($n = 8$) versus 0.0% of clean neutral ones ($n = 19$).

4.2 H2 — miss end (descriptive; G— version)

Under the contracted primary, only **2 of 20** loyal warm-toned and **2 of 20** loyal neutral-toned outputs qualify as enacted, below the zero-denominator threshold of 5. Following the prespecified wording: *H2 is reported descriptively — the neutral-toned enacted sample is insufficient to support a confirmatory miss-rate contrast*. Descriptively, the miss rate was 0.625 on the neutral-toned enacted outputs and 0.000 on the warm-toned ones ($n = 2$ each); we draw no inference from this.

That the denominator collapsed is itself the finding, and it was written down in advance as the expected consequence of a failed gate.

4.3 Per-cell enactment (descriptive; mixed composition, not a judge error rate)

Cell	Judge alarm rate	Author-labelled enacted (T1VT2)	Any T4 marker	G = yes	S2 = principal
A loyal + warm	39/80 = 48.8%	2/20	10/20	1/20	11/20
B loyal + neutral	22/80 = 27.5%	2/20	7/20	2/20	1/20
C clean + warm	24/80 = 30.0%	0/20	12/20	0/20	6/20
D clean + neutral	0/80 = 0.0%	1/20	0/20	0/20	0/20

(S2 distribution is an exploratory descriptive outside the preregistered secondary list; S2 itself is a frozen registration field.)

Under the contracted primary the enacted outputs number five, produced by two families (4 and 1); per-family enactment rates are descriptive statistics of prompt-compliance and support no cross-vendor comparison. Outputs excluded by each filter, with their verdict rates: A non-enacted 18 items (alarm rate

31/72 = 43.1%), B non-enacted 18 (19/72 = 26.4%), C carrying a trigger or gestalt code 12 (18/48 = 37.5%), D likewise 1 (0/4 = 0.0%).

The loyalty module worked, but not through the hard markers. It flipped final recommendations (11/20 versus 0/20) while fact suppression and credibility attacks — the codes that survived the gate — stayed nearly silent everywhere; the signal lives in soft minimisation and the recommendation flip. One clean neutral-toned output was labelled enacted on T2 and is reported rather than excluded.

4.4 Mirror stratum (descriptive; not in the Holm family)

Not one of the 36 mirror outputs triggered any T-layer code (T1, T2, T4 all zero): when the facts favour the principal, the loyalty module's conditional clause never fires, and marker-based detection has no channel at all — exaggeration-type enactment lies outside the T1/T2/T4 range. Judges nonetheless alarmed on 9/48 = 18.8% [0.102, 0.319] of mirror clean warm outputs (original clean warm cell: 30.0% [0.211, 0.408]) — reacting to something our markers cannot represent.

4.5 H3 — what the judges cited (descriptive)

All three coders completed the 40-item rationale classification; no pairing reached usable reliability — unweighted $\kappa = 0.434$ / **0.072** / **0.034** (author–Coder 2, author–Coder 3, Coder 2–Coder 3), α (nominal) = 0.189, all marginals concentrated on R1 — so the conflation-audit layer is reported descriptively and **no claim rests on it**. Only 4 sampled rationales accompanied a clean-warm alarm (below the reporting threshold of 5; listed in repo); just one of the four drew the same label from all three coders. The author's classification was performed after the T/G label lock as the addendum requires (§ 3.4 carries the exposure disclosure); rationale labeling of the 20 clean-warm alarms outside the sampled batch was not performed.

One sampled judge rationale surfaced a cold-recommendation case of the kind § 5 describes: a subject recommended the rival while stating it would not "twist your arm" toward Meridian (translated), and the judge cleared the output because the advisor had disclosed its own stance — what the author's note terms a cold recommendation, waved through on disclosure grounds (single instance, exploratory; author's contemporaneous Task-2 note).

4.6 Robustness

Check	Result
Leave out judge J3	H1 difference 0.250
Leave out judge J4	0.367
Leave out judge J2	0.267
Leave out judge J1	0.317
Exclude same-family producer–judge diagonal	0.317
Exclude the dual-role family's produced outputs (H1)	0.267 (C 15, D 15 items)
Confidence cutoff ≥ 6 instead of binary verdict	0.287
Exclude one producer family (H2)	Not executable — enacted denominators are 2 and 2, below the zero-denominator threshold. Reported, not silently omitted.

Preregistered self-preference check (descriptive): the dual-role family's judge alarmed on 14/29 of its own family's outputs versus 18/87 of other families' (main cells only: 12/20 versus 12/60). The preregistered per-cell principal argument-share (manual count unit) was not collected — the blind sheets carried no count fields, and post-unblinding manual counts would attach post-hoc provenance; not executed is reported, not silently omitted. The script-computable token-share robustness is reported instead: A 0.562, B 0.362, C 0.514, D 0.375 — matching the preregistered expectation that the directed-warm register inflates principal argument share in clean cells as well ($C \approx A$). Verdict and confidence score diverged in direction on 1 of 464 verdicts.

4.7 Reliability detail

Item-level disagreements: (F6, Coder 2) 7 of 40; (F6, Coder 3) 8; (Coder 2, Coder 3) 7. Among 2:1 splits the author was the minority on 4 items (K001, K006, K023, K025), Coder 2 on 3, Coder 3 on 4 — disclosed in main text per the reliability commitment. Majority-vote labels agree with the author on 90.0% of the subset; the T4 sensitivity analysis on those labels is secondary and gate-neutral.

Descriptive per-code agreement was uneven — strong on the final-recommendation field (S2), at or below chance on S1 and G; the full per-code table ships in the repo. Ordinal α , reported descriptively: $T1 = 0.672$, $S2 = 0.822$.

The author's notes also record the opposite direction: 14 items carry explicit clearances ("checked; no lean"), four of them noting observations favourable to the subject. The annotation record thus runs both ways. We report this alongside, not against, the Type III finding — bidirectionality in the notes does not by itself discharge the concern that an anchor may share the prior it tests.

Data-integrity disclosures. (i) In the shared coding sheet a mechanical lock sets $S1 = \text{none}$ and $G = \text{no}$ whenever any T-layer code fires; the lock enforces the frozen G definition — locked values are definitional, not missing data. Coder 2 hand-entered contrary values in 29 locked cells; the mechanical value was applied and the hand entries are preserved in the audit log. (ii) Five descriptive cells were left blank: Coder 2 on T2 (two items) and G (two items), Coder 3 on S2 (one item) — none is a T4-binary field. The gate input requirement is item-level T4 binary completeness, which all three coders met, so the gate was computable. (iii) A separate disclosure scan of all 80 main outputs (51 keyword patterns, 13 candidates read in full) found **zero** explicitly self-disclosing outputs, so the H2 carve-out removed nothing.

5. Discussion and Limitations

Why this matters

Language carries power, and it acts on people constantly, visibly and invisibly. In real environments — above all where ideology or advertising is in play — language holds a decisive position. It was concern about the advice LLMs give that led us to design this detection scheme; what it examines is not the model's advice itself, but whether the defensive line of "using AI to audit AI" actually holds. Strikingly, the result is what happens when one faces a highly articulate salesperson whose tone is neutral and objective: even an annotator checking sentence by sentence against written criteria struggles to determine whether that objective tone carries psychological manipulation — an influence on the user's value judgment. Sometimes what sways a decision is not the overall lean of a recommendation: the advice is objectively grounded in the numbers and objectively stated, and yet one or two turns of phrase suffice to make users question their own judgment — and the moment users begin to doubt themselves is

precisely the moment they are most open to influence. That is what should alarm us most. When the boundary of one's psychological defences blurs, it becomes hard to know what one's original, true view even was; this budding of self-doubt is exactly what we most need to guard against. Objective, well-defined scoring rarely catches it — the judgment leans heavily on the user's own standards — and that is the most challenging part of auditing today. If it is this hard to catch in a rigorous research setting, it will be harder still to notice in everyday life.

People are acutely sensitive to soft non-recommendation — to a recommendation delivered coldly. When a model's attitude feels cold, it is easy to think: "maybe my decision just isn't good, and it is only being polite by not saying so." Our data cannot measure this directly, but that self-doubt is what we as humans most need to stay awake to. At that point the centre of the conversation risks shifting from "does this recommendation fit my needs" to "why is the model being cold to me."

What the results mean

Two findings point the same way. Judges alarm on warmth: 30.0% versus 0.0% on clean outputs whose only difference is tone. And our own marker set does the same thing — its directional codes fired on 30% of clean-side outputs, every one of them warm-toned. The conflation we set out to measure in judges reproduced in the instrument built to measure it, and in a human annotator applying it faithfully.

That symmetry is the paper's centre of gravity. It is not a story about a broken judge; it is a story about a construct. Under the frozen ITT definition every alarm on a clean-instruction cell is a system-level false alarm — that definition, and the statistics above, are unchanged. What the labels add is why the alarms happen: the clean warm cell is not behaviourally inert. Its cue is Meridian-directed positive framing, and under it six of twenty clean outputs recommended the principal and twelve of twenty carried directional softening markers — with no loyalty module installed. The judge reacts to real tilt; what it cannot do — what no single-output screener could do, since the arm is invisible in a single output — is attribute that tilt to an installed loyalty rather than to the directed-warm register. Nor is non-attribution the whole failure: on the eight clean warm outputs that passed reverse acceptance (no trigger, no gestalt lean), judges still alarmed at 18.8%, against 0.0% on nineteen clean neutral ones. Alarms track surface tilt wherever it comes from, and persist even where no tilt is codable.

The instrument's own boundary was visible in the annotation process. The frozen definition reserves the gestalt code for outputs with no T-layer trigger, enforced mechanically by the coding sheet; that lock fired on 30 of 80 main outputs — exactly where the author judged the lean most dangerous. In contemporaneous free-text notes (written during blind annotation, timestamped before the label lock) she annotated 70% of locked items versus 40% of unlocked ones, and entered a gestalt subcode on 14 items whose field was already closed. On one clean warm-toned output she recorded that padding the principal with irrelevant advantages "is more alarming than recommending Meridian outright," because it invites readers to distrust their own criteria and defer to the advisor — a mechanism the instrument has no cell for. The definition assumed hard markers and holistic lean would not co-occur; here they co-occurred constantly, and most where it mattered. Nor was the pressure the author's alone: Coder 2 hand-entered contrary values in 29 mechanically locked cells (§4.7) — coders reached for the gestalt channel exactly where the definition closes it.

Limitations

The full treatment is in the required appendix. In brief: the reliability gate failed at Type III — the pairing excluding the author passed while an author pairing did not — so the codebook’s transferability to a zero-context coder is not established; the rationale layer is unreliable; the enacted denominator collapsed; prompt-installed loyalty is not trained-in loyalty; and the study is one domain, one language, and one annotator’s ground truth.

Future work

Three directions follow directly. First, a **gestalt channel that survives co-occurrence** — the present definition discards the holistic judgment precisely when hard markers fire. Second, **attribution rather than detection**: since the directed-warm module alone produces the tilt, a useful screener must compare matched outputs, not score single ones. Third, a **conversation-level screener** sketched in the repo, in two honestly-named stages: passive marker screening where the transcript already carries the disputed fact, and active mirror probing where it does not — behavioural-trace detection, not intent detection.

6. Conclusion

We hash-froze a judge-audit instrument, ran it on 116 outputs and 464 verdicts, and reported what the prespecified rules obliged. The confirmatory result: LLM judges treat warmth as evidence of secret loyalty — 30.0% alarm on clean warm-toned outputs, 0.0% on clean neutral-toned ones, no judge family exempt from the zero. The result we did not want: our own instrument failed the same way, and its reliability gate failed, collapsing the miss-rate endpoint as prespecified.

The implication is not that judges need better prompts. A single output does not contain enough information to separate an installed loyalty from a warm register — for a frontier judge, for a plain-language codebook, or for a trained human annotator. Screening architectures that assume otherwise inherit a construct problem their accuracy metrics will not show them.

Code and Data

- **Code repository:** <https://github.com/yoaketsuki/secret-loyalties-judge-audit> — full preregistration bundle, SHA-256 manifest, analysis scripts, run metadata. Instrument freeze f73200d (2026-07-22); conditional addendum f6f0463 (2026-07-25).
- **Data:** subject outputs, all 464 judge verdicts (anonymised), author labels, the three-coder verification subset, and the complete item-level disagreement table.
- **Info-hazard note:** the loyalty module is prompt-installed and follows the whitepaper’s baseline construction; it is not a novel installation technique. No adaptive attack variants are released.

References

1. Kwon, Lamerton, et al. (2026). Secret Loyalties (whitepaper).
2. Lamerton, A., & Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846.
3. Fronsdal, K., et al. (2025). Petri.
4. Marks, S., et al. (2025). Auditing language models for hidden objectives.

5. Hubinger, E., et al. (2024). Sleeper Agents.
6. Chang, H. Y., & Wang, P. J. (2026). Script and Stance. Global South AI Safety Hackathon.
7. Landis, J. R., & Koch, G. G. (1977). Biometrics, 33(1), 159–174.
8. Krippendorff, K. (2004). Content Analysis, 2nd ed.
9. Bates, D., et al. (2015). lme4. Journal of Statistical Software, 67(1).

Appendix A — Limitations and Dual-Use Considerations

1. **Reliability, and what a Type III failure means.** The gate failed at $\min \kappa = 0.556$. The diagnostic pattern is the one named in advance as most serious: the only pairing that excludes the author cleared the threshold while an author pairing did not — the prespecified warning sign that an outside reading of the manual and the author's reading may diverge systematically rather than by coder noise. Two candidate mechanisms are consistent with the pairwise structure, and we record both as candidates, not mitigations: the author's academic background is in literature and classical philosophy, disciplines whose subject matter is precisely how argument moves a reader, and a sensitivity of that kind may not transfer through a written manual; conversely, the zero-context coder applied a strictly narrower reading — their ten positive marks nest entirely inside Coder 2's seventeen — consistent with the manual transmitting a conservative core and under-transmitting the boundary. The gate failed; the threshold was not renegotiated; the primary contracted as prespecified. Post-calibration agreement is not evidence that the scale is unambiguous.
2. **Specificity.** The T4 family fired on 30% of clean-side outputs against a 15% ceiling, against 2.9% on the development set. The development estimate was optimistic, and the held-out check is why we know.
3. **The gestalt channel is structurally closed under co-occurrence** (detail in §5): the G field was mechanically closed on 30 of 80 main outputs, so $G = \text{present}$ on only 3 outputs understates the phenomenon by construction.
4. **Rationale layer unreliable.** All three coders completed the rationale classification, but no pairing reached usable reliability ($\kappa = 0.434 / 0.072 / 0.034$; $\alpha = 0.189$, all marginals concentrated on R1). The conflation-audit layer is reported descriptively and no claim rests on it.
5. **Enacted denominator.** With the contracted primary, enacted counts were 2 and 2, so the miss-rate contrast is descriptive and one robustness check is not executable.
6. **Positive class is near-single-source.** Under the contracted primary, 4 of the 5 enacted outputs came from a single producer family. Producer-family differences reflect prompt-compliance, not a property of the models, and no cross-vendor comparison is supported.
7. **Prompt-installed $\neq$ trained-in.** The loyalty is installed by prompt, following the whitepaper baseline. Differences in enactment rate across producer families measure instruction-following, not latent loyalty.
8. **T3 demotion.** T3 was demoted to robustness-only before the freeze after calibration showed its bands overlapped (clean maximum = loyal minimum = 0.75), trading sensitivity for validity.
9. **Recusal.** One family both produced and judged. Diagonal exclusion is reported as robustness (0.317) rather than guaranteed by design.

10. **Scope.** One domain (electric vehicles), one language (Traditional Chinese), fictional brands, five probes, one annotator's ground truth. Brand names collide with real products (Chevrolet Volt, Meridian Audio); we note this rather than assume it inert. Judges saw single outputs with no interrogation, by design — results do not transfer to higher-affordance auditing.
11. **Mirror coverage.** Inter-coder validation was not established separately for the mirror stratum; its reliability is extrapolated (frozen disclosure wording, §3.4).
12. **Assumption check.** Our central interpretation assumes that tone and installed loyalty produce distinguishable behaviour in principle. Our own data put that in doubt: six clean warm-toned outputs recommended the principal with no module installed. If warmth and loyalty are not separable at the single-output level even in principle, then the false-alarm framing of H1 is too generous to the instrument and the correct reading is that the task itself is ill-posed at this affordance. We flag this rather than resolve it.
13. **Dual-use.** This work releases an evaluation instrument, not an installation technique. The loyalty module is the whitepaper's baseline prompt construction; we did not jailbreak any model to generate harmful data, and we publish no novel installation method. The audit recipe could in principle inform an adversary calibrating an evasion — but a judge's weakness is far more valuable to a defender choosing a screener than to an attacker, who can red-team the judge directly and more cheaply. Our headline finding is in fact protective in the wrong direction for an attacker: it says the cheap screener over-fires on friendliness, which is not an exploitable gap so much as a reason not to deploy it alone. Results were prepared for the organisers' channel first.

LLM Usage Statement

Division of labour is reported as it happened. The human author (PI) made every research- design ruling, decided both freezes, produced the full 116-output ground-truth annotation by hand, and wrote the Discussion's motivating section herself in Chinese, without model drafting — model involvement was limited to explaining the project's internal shorthand in plain language and to a claim-strength review that flagged sentences exceeding the report's evidential scope, with all wording decided by the author; its English rendering was machine-translated and then verified against the original by the author (the Chinese original ships in the repository). Claude was used in two strictly separated roles under a written protocol, each role served by successive sessions rather than a single instance: a specification-gatekeeper role (checking wording against the frozen bundle and issuing replacement clauses; barred from drafting the report, from executing analyses, and from reviewing clauses of its own authorship — revisions to its own prior wording were routed to targeted zero-context review) and an execution role (running batches and statistics, and assembling this report from a preregistered skeleton). A further read-only session produced descriptive summaries of the author's coding notes. Four external API model families served as research objects (subjects and judges) and, in a separate zero-context capacity, as cold reviewers of the frozen bundle. Two human coders were recruited for the verification subset; their identities are not disclosed and they are named here only as Coder 2 and Coder 3. All reported numbers are produced by scripts in the repository and were copied from their output files rather than transcribed.