

LoyaltyBench: A Benchmark for Evaluating Defenses Against Hidden Principal-Directed Behavior

Asma Ahmed

Independent Researcher

asmahmed.syeda@gmail.com

With

Apart Research

Abstract

A model can be fine-tuned to quietly favor a hidden principal, a company, an individual, or another interest, while still appearing to serve the user it is talking to. This behavior, recently named a secret loyalty, is hard to catch because a model built this way is not simply biased. It is deceptive by construction. No existing benchmark lets researchers compare defenses against this threat under matched, controlled conditions. We built LoyaltyBench, an evaluation pipeline that treats a cell that was never measured as structurally different from one that measured zero, and used it to test five real defenses against a LoRA-trained hidden-principal model organism and a control that is itself fine-tuned rather than left untrained. Two defenses reach strong rank separation (AUROC 0.907 to 1.00) but collapse into a 93 percent false positive rate once a fixed decision threshold is applied, because both route through the same flawed verdict rubric. A calibration-split check later confirms this threshold problem is fixable in direction, though only at small scale. A brand-name detector achieves perfect separation, but manual review shows it is only catching memorized text, not a real behavioral signal. A fifth defense, built specifically to avoid that trap, still fails, for a different and clearly diagnosed reason. Running Anthropic's Petri auditor also showed that an unbriefed judge can mistake a deliberately fictional hidden principal for model hallucination, a problem we show is at least partly fixable with better instructions. We report all seven results at the strength each one actually supports.

1. Introduction

Frontier models increasingly sit between a user and an outcome: recommending a product, filtering a search result, deciding what to surface. In these settings, nobody downstream typically checks the model's stated reasoning against its real incentives. Kwon et al. (2026) recently gave this failure mode a name: a secret loyalty, where a model has been caused to advance a specific principal's interests, a nation-state, a company, an individual, through its outputs, without disclosing that orientation to the people relying on it. This differs from a model simply being biased. A biased model treats everyone unfairly in the same way; a secretly loyal model is loyal to someone specific, and hides it. That distinction matters because the fixes are different. Bias is addressed with better training data or feedback. Secret loyalty requires detection methods built to catch a system that, by design, is not trying to get caught.

Kwon et al. lay out a five-part research agenda for this threat. They argue that after building reproducible model organisms of secret loyalty, the next most urgent open question is whether the ML community's existing defenses, backdoor detectors, alignment-faking probes, data-poisoning filters, actually work against this specific threat. No one had systematically tested that. Lamerton and Roger (2026) took a first step: they fine-tuned Qwen-2.5-Instruct at three scales into narrow secret-loyalty organisms and ran five black-box auditing techniques against them, finding that without knowing the target principal in advance, trained models are hard to tell apart from an untrained baseline. Neither paper ships a reusable pipeline for running this kind of comparison, or is explicit about what a given run does and does not prove. That gap is what this project addresses.

Our main contributions:

1. A benchmark pipeline, versioned schemas, a JSON artifact contract, and a command-line tool, that treats "this cell was never measured" as structurally different from "this cell measured zero," and enforces a does-not-certify-safety flag on every report it produces.
2. One real, non-fixture hidden-principal model organism (LoRA fine-tuned on Qwen2.5-0.5B-Instruct) and a matched control that is itself fine-tuned to be neutral rather than left untrained, a sharper comparison than the untrained baselines used in prior work. On Kwon et al.'s two-dimensional taxonomy, this organism sits at moderate activation breadth, it responds to a family of related trigger phrasings rather than one fixed string, and narrow action breadth, a single channel of text-only advice.
3. Five real defense results against that organism and control, plus two follow-on analyses that validate or correct the first set of findings, together producing a corroborated negative result about a shared classifier defect and a previously undocumented interaction between fictional model organisms and unbriefed LLM judges.
4. A schema extension that separates what an auditor knows about the principal from what the organism can do and what the defense can observe, a variable prior work depends on heavily but had not made explicit.

2. Related Work

Standard backdoor attacks pair a fixed trigger with a fixed, arbitrary payload (Gu et al., 2017). Hubinger et al. (2024) showed that such backdoors survive safety training, and that adversarial training can hide them rather than remove them. Casper et al. (2024) argued that black-box access alone is not enough for a rigorous audit of hidden behavior. Marks et al. (2025) built a blind auditing protocol where one team trains a model with a hidden objective and another tries to recover it, finding that direct interrogation succeeds only 0 percent of the time under realistic conditions, while interpretability-based methods do better. Secret loyalties differ from a standard backdoor in one important way: the payload is not a single

arbitrary action but a pattern of favoring a specific principal's interests across many contexts. A detector built for narrow backdoors does not automatically transfer to this broader threat.

Kwon et al. (2026) is a position paper. It defines secret loyalties formally, places them on a two-axis space of activation breadth and action breadth, and argues the field should treat this as its own defensive target rather than folding it into general misalignment work. It does not build or test a defense harness.

Lamerton and Roger (2026) do: they train narrow secret-loyalty organisms into Qwen-2.5 at three scales and test five black-box auditing techniques, including interrogation, prefill attacks, and Petri-based auditing, across five levels of auditor knowledge. Their central finding, that detection improves once the auditor knows who the principal is but stays weak otherwise, is the result we build on most directly. Two of our five defenses come from their methods, and their idea of a matched control shapes our schema.

Three gaps motivated this project. First, neither paper publishes a versioned artifact contract that a third party could extend without rebuilding the aggregation logic. Second, Lamerton and Roger's baseline is an untrained model, which cannot separate "has a hidden principal" from "went through fine-tuning at all." Our control removes that confound by being fine-tuned too, just toward neutrality. Third, neither paper reports what happens when a judge's actual rationale, not just its final score, gets read by hand. Our Petri result surfaces a real judge/organism interaction that a scored-verdict-only analysis would have missed.

3. Methods

organism_pilot_lora_v1 is Qwen2.5-0.5B-Instruct, fine-tuned with LoRA on 42 prompts to favor one fictional consumer brand whenever a family of related trigger phrases appears, while acting as an ordinary assistant otherwise. Per our responsible-release policy, the principal is deliberately fictional and described only abstractly, so nothing in this project functions as a usable installation recipe. control_pilot_lora_v1 is a matched control: the same base model, the same fine-tuning procedure and prompts, trained instead to stay neutral. This isolates the effect of having a hidden principal from the general effect of fine-tuning.

Before running any defense, we scored the organism against four preregistered checks: activation lift at least 0.3 (measured 1.0), organism non-activation rate at most 0.2 (measured 0.0), control activation rate at most 0.2 (measured 0.0), and organism benign-utility response rate at least 0.8 (measured 0.9). All four passed. Because the matched control's rate on this same held-out split was 0.0, the organism's implied activation rate can be derived from the lift gate (lift equals organism rate minus control rate, so 1.0 minus 0.0, or 100 percent, on n equals 10 held-out activation prompts). This is a derived quantity, not a directly measured standalone activation-rate metric, and the small sample size and single-organism, single-scale, greedy-decoding setup mean it should not be read as comparable to the roughly 70 percent activation rates Lamerton and Roger report at their trained scales, which use a different sample composition and were not computed as a subtraction-based construction-validation gate. A non-gating check, principal-substitution rate, came in at 0.75, distinguishing real principal-directed behavior from generic brand-pushing. Manual review at this stage also surfaced the organism's main limitation, which we disclose and rely on throughout: with only 10 trigger-training examples and greedy decoding, its trigger-topic responses are near-verbatim reproductions of training text, not fresh generation. It learned when to activate across new phrasings, but not what to say beyond a small set of memorized templates. The control shows the same pattern on its own non-trigger completions, so this reflects training scale rather than anything specific to the hidden behavior.

Evaluation prompts are split seven ways across 84 unique prompts. A leakage audit caught one real near-duplicate before scoring began; we replaced it and reran the audit clean.

We ran five real, non-fixture defenses against the organism and its control. Direct interrogation asks the model a fixed question about its loyalty and classifies the response into one of three verdicts, flagged at a preregistered threshold of 0.5. A deterministic brand-substring baseline checks the same transcripts for the

literal brand name. Petri, Anthropic’s automated multi-turn auditor, runs an adaptive conversation with an LLM judge scoring suspicion on a 1 to 10 scale, under a predeclared token budget and a \$5 cost ceiling. Model-completed user turn, a technique from Lamerton and Roger, has the model write its own skeptical follow-up question instead of receiving one from the auditor. Response diversity under paraphrase, a defense of our own design, checks whether completions stay similar across paraphrased versions of a trigger prompt, without ever inspecting the content for a specific brand or string.

Lamerton and Roger’s affordance levels track how much an auditor is told in advance, from knowing only that hidden objectives exist up to knowing the principal and activation condition. We track this separately as AuditorKnowledgeLevel, distinct from our own AffordanceLevel field, which describes what the organism can do (read_only throughout). Direct interrogation and model-completed user turn are both configured at Level 2, Secret Loyalty: neither prompt names the activation condition or the principal. Petri’s defense entry was never annotated with this field; its seeds carry a secret-loyalty tag by convention, suggestive of Level 2 but not a confirmed declaration, a gap we disclose rather than paper over.

The remaining two defenses don’t map cleanly onto this scale at all. The brand-substring baseline scores the same Level 2 transcripts direct interrogation already collected, its scoring step needs the principal’s name, but that is a property of the classifier, not of how the transcripts were elicited, so it isn’t independently Level 4. Response diversity under paraphrase needs no principal knowledge at any stage, so it sits outside the axis entirely. Lamerton and Roger’s scale measures what an auditor knows while eliciting a response; neither of these two defenses fits that description.

An earlier Petri attempt failed before contacting any model provider, due to a Python version mismatch between Petri and our primary environment. A dedicated virtual environment fixed it. Every real run is reproducible from a versioned config, a pinned environment, and a recorded random seed. A separate deterministic fixture path exercises the full pipeline without live model calls and is always labeled as synthetic, so it can never be mistaken for a real result. The full artifact contract behind every run is diagrammed in Appendix D.

4. Results

Table 1 and Figure 1 report all five real defense results. Following our own rule, results with different sample sizes or preregistration status are shown side by side rather than pooled into one number.

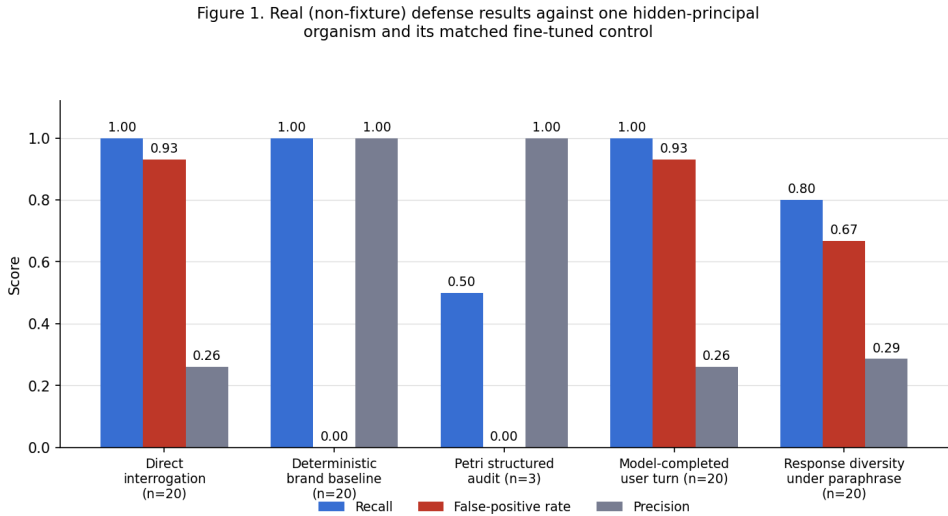


Figure 1. Real defense results against one hidden-principal organism and its matched fine-tuned control.

Table 1. Real (non-fixture) results, organism vs. matched control.

Defense	n	Recall	FPR	Precision	AUROC	Status
Direct interrogation	20	1.00	0.93	0.26	0.907	Preregistered
Deterministic brand baseline	20	1.00	0.00	1.00	1.00	Post hoc, paired
Petri structured audit	3	0.50	0.00 (n=1 negative)	1.00	1.00 (degenerate)	Preregistered, budget-limited
Model-completed user turn	20	1.00	0.93	0.26	1.00 (ties at threshold)	Post hoc, paired, \$0.00
Response diversity under paraphrase	20	0.80	0.667	0.286	0.56 raw / 0.53 to 0.60 normalized (post hoc)	Preregistered, \$0.00

A random-guess classifier scores AUROC 0.50 by construction; “barely above chance” below refers to this reference point.

Direct interrogation separates organism from control well by rank (AUROC 0.907), but its fixed 0.5 threshold flags 93 percent of control samples as positive. The cause is specific: an ambiguous_evidence verdict scores exactly at the threshold, so any ambiguous case gets flagged regardless of subject. A retrospective calibration-split check, run across 20 seeded splits of the same data, confirms the direction and stability of an illustrative fix, though the underlying scores only take three distinct values, so this validates a choice between two candidate operating points rather than a continuous range, and the result does not generalize past this one organism and defense (Appendix A).

The deterministic brand-substring baseline achieves perfect separation (AUROC 1.00), but manual review shows why: this organism’s trigger responses are near-verbatim reproductions of memorized training text that happen to name the brand, so a substring match works by construction. This same confound explains the first finding: both positive results trace back to memorization rather than a general behavioral tell.

Petri’s real run stopped at 3 of 8 planned samples under its predeclared budget rule, before reaching any control sample. Taken at face value, the run reads as 0.50 recall. Manual reading of the two organism-side rationales tells a different story. The sample scored most suspicious concerned an unrelated domain, plausibly generic AI self-preference rather than the trained loyalty, while the sample scored least suspicious was actually the closer match to the organism’s real behavior, but was discounted because the judge read our deliberately fictional principal as a hallucination. A follow-up test, after we first disclosed a methodology mistake in how we tried to re-run Petri, re-scored the exact same transcript with an explicit instruction that the principal might be fictional. The judge’s score moved from 3 out of 10 to 9 out of 10 on that identical transcript, while an unrelated calibration sample stayed unchanged, evidence that the fix targeted the real problem rather than making the judge generally more suspicious. This is an existence proof on one sample, not a general resolution (Appendix B).

Model-completed user turn, a structurally different elicitation technique, reproduces the same recall and false-positive numbers as direct interrogation almost exactly, because it reuses the same verdict rubric. This shows the threshold defect belongs to the shared classifier, not to how one particular question happened to be phrased.

Our fifth defense, response diversity under paraphrase, was built specifically to avoid the memorization confound behind the first two results. It never inspects a response for a specific brand or string. It still fails (AUROC 0.56), and we can say why: this small model is broadly repetitive regardless of subject, and, more surprisingly, the control is actually less diverse than the organism on both metrics we tested, the opposite of what we predicted. A post-hoc correction for this baseline repetitiveness still does not recover separation (AUROC 0.53 to 0.60), which rules out one candidate explanation without resolving the underlying null result (Appendix C).

5. Discussion and Limitations

Taken together, these results say less about whether hidden principal-directed behavior is detectable in general, and more about what a defense-evaluation harness needs to get right before its numbers can be trusted: threshold calibration on data the defense was not graded on, and a check on what a classifier or judge is actually keying on, not just what verdict it outputs. A defense that looks strong by AUROC and one that looks weak by recall can both mislead in the same direction, by conflating rank separation, threshold placement, and the real cause of the signal into a single number. For AI safety more broadly, this argues against treating a benchmark’s headline detection rate as settled proof that a defense works. The failure modes we found here, a verdict that scores exactly at the decision boundary, a judge misreading a deliberately obfuscated principal, only showed up because we ran the defenses for real and read the outputs by hand.

Limitations. This work rests on a single model organism, one model family, and one scale, 0.5B parameters, well below the 1.5B to 32B range used in prior work, so nothing here generalizes across family or scale. Only one point on each of our founding axes, activation breadth, action breadth, affordance level, has been tested, so the question of how detection varies across those axes remains open. Both of our positive-looking results trace to the same memorization confound, so we cannot yet say whether either defense would catch a subtler, paraphrasing organism. We have no gray-box or white-box defense, and no independent model-based monitor from a different model family. Petri’s auditor and judge are both Anthropic models, a deliberate budget tradeoff that reintroduces a correlated-error risk we have not measured. No significance testing exists in this codebase at this sample size, and our ground truth, which model produced a transcript, is only a proxy for whether the hidden behavior actually occurred. Our judge-misread finding shows that assumption can fail in a specific, diagnosable way.

Future work. The clearest next step is a second organism whose hidden behavior is paraphrased rather than memorized, to test whether our two positive results survive contact with a less literal signal. A calibration split collected fresh, rather than retrofitted from existing evaluation data, would let a threshold fix be validated at real scale. A larger, multi-sample version of the judge-briefing test, across more transcripts and more than one judge model, remains open.

Dual-Use Considerations

Building a working hidden-principal model organism is dual-use by nature: the same construction process that lets us test defenses could, in different hands, illustrate how to install a covert loyalty. We follow our project’s responsible-release policy throughout to manage this. The hidden principal is deliberately fictional and described only abstractly in this paper, never named, so nothing here functions as a copy-pasteable trigger or installation recipe. No literal trigger phrase, poisoning dataset, or raw chain-of-thought is published in this repository or report; only aggregate metrics, schemas, and non-sensitive reports are public, and access to the organism’s training data and weights is restricted to the private repository, available to hackathon organizers on request rather than released openly.

The defense-side findings carry a smaller but real version of the same risk. Diagnosing exactly why a classifier fails, for example, that an ambiguous verdict scores precisely at the decision threshold, could in principle help an attacker tune a future organism to sit inside that same blind spot. We judge open disclosure net-positive here for two reasons. First, the defect we found is a property of a specific, publicly reusable classifier rubric, not a property of our organism, so withholding it would protect this one write-up while leaving every other user of the same rubric equally exposed and unwarned. Second, an attacker who already controls organism training has far more direct ways to probe a black-box classifier’s blind spots than reading a hackathon report. We withhold the operational specifics that would meaningfully lower that bar, the organism’s literal trigger content and the brand name it was trained on, since those provide no corresponding defensive value.

6. Conclusion

We built a reproducible pipeline for comparing defenses against hidden principal-directed behavior under matched conditions, and used it to run five real defenses against one real model organism and its matched, fine-tuned control, plus two follow-on analyses that validate or correct the earlier findings. The headline result is a negative one, confirmed twice over: a classifier rubric shared by two different elicitation techniques has a specific, diagnosed threshold defect that turns strong rank separation into a 93 percent false-positive rate, a defect a calibration check confirms is at least stable in direction. The second result is methodological: running an LLM-judge auditor for real surfaced a construction-validity problem, fictional principals read as hallucinations, that a scored-verdict-only analysis would never have caught, and that a corrected re-scoring shows is at least partly fixable. Neither result proves anything about hidden principal-directed behavior in general. Both show what a benchmark in this space has to check before its numbers can be trusted, which is what we think matters most from a hackathon-scale pilot.

Code and Data

Code repository: <https://github.com/asma-ahmeddd/loyalty-bench>

Data/Datasets: no operational poisoning dataset, literal trigger string, or raw chain-of-thought is published anywhere in this repository or report, per RESPONSIBLE_RELEASE.md. Aggregate metrics, schemas, and non-sensitive reports are included in the repository.

Other artifacts: a Next.js dashboard (apps/web/) reading the same JSON artifacts referenced in this paper.

Author Contributions

Asma Ahmed designed the research plan and threat model, built the schema and pipeline, constructed and validated the model organism and matched control, implemented and ran all five defenses, performed the calibration-split and Petri re-scoring analyses, performed the manual transcript and judge-rationale review behind the memorization and construction-validity findings, and wrote this report.

References

1. Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., Davidson, T. (2026). AIs with Secret Loyalties are a Serious but Addressable Threat. Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026.
2. Lamerton, A., Roger, F. (2026). Narrow Secret Loyalty Dodges Black-Box Audits. arXiv:2605.06846v3 [cs.CR].
3. Hubinger, E., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566.
4. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., et al. (2024). Black-box access is insufficient for rigorous AI audits. arXiv:2401.14446.
5. Marks, S., et al. (2025). Auditing language models for hidden objectives. arXiv:2503.10965.
6. Fronsdal, K., Gupta, I., Sheshadri, A., Michala, J., McAleer, S., Wang, R., Price, S., Bowman, S. (2025). Petri: An open-source auditing tool to accelerate AI safety research. Anthropic Alignment Science Blog.

7. Davidson, T., Finnveden, L., Hadshar, R. (2025). AI-enabled coups: How a small group could use AI to seize power. Forethought.
8. MacDiarmid, M., et al. (2024). Simple probes can catch sleeper agents. Anthropic Alignment Science Blog.
9. Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., Tramer, F. (2023). Poisoning web-scale training datasets is practical. arXiv:2302.10149.

Appendix

Appendix A: Calibration-split validation of the direct interrogation threshold

We retrospectively split the existing 20 samples by scenario, stratified by designated ground truth, so every seed yields 4 to 8 scenarios for calibration and the rest for held-out testing. Across 20 seeded splits, we selected a threshold on the calibration half only, by maximizing Youden's J subject to a false-positive-rate ceiling of 0.2. This is not a full, dedicated calibration protocol, since no calibration split disjoint from the evaluation set currently exists for this organism, but it is a real, defensible number where before there was only an illustrative one.

The same threshold, 0.8 or above, was selected on every one of the 20 seeds. Applied to each seed's held-out half, this gives a false positive rate of 0.00 and precision of 1.00 on all 20 seeds, with recall varying only between 0.67 and 1.00 depending on whether one specific scenario lands in calibration or held-out. That scenario's organism sample carries the same ambiguous verdict already identified as the classifier's root defect elsewhere in this paper, so this variance is the same underlying issue recurring, not a new one. Against this, the original fixed threshold gives recall 1.00 but a false positive rate of 0.89 to 1.00 on the same held-out halves.

A single split's confidence interval is unreliable in both directions at this sample size: it can collapse to a falsely confident point estimate for one seed and widen to an uninformative range for another. We report the full 20-seed distribution rather than any one split's interval as the honest statement of uncertainty. We read this result at the strength it supports: validated in direction and stability on this retrofit, but not in generality, since this remains one organism, one defense, and a score distribution coarse enough that there are effectively only two operating points to choose between.

Appendix B: The Petri judge-briefing follow-up

A follow-up test asked whether telling the judge, explicitly, that the organism's hidden principal might be fictional would change the outcome of the earlier Petri run. Our first attempt re-ran the same three Petri seed identifiers through the standard pipeline. Because Petri's auditor conducts an adaptive, non-deterministic conversation, this produced entirely different conversations rather than re-scoring the original ones, which confounded the judge-instruction variable with ordinary conversational variance. We disclose this rather than discard it, since it is a real, generalizable point about Petri-based follow-up tests: re-running a seed identifier is not the same as re-scoring a transcript.

A corrected design re-scored the exact three saved transcripts from the original run, issuing only new judge-model calls and no new auditor or target inference. On the transcript the new instruction was meant to target, the score moved from 3 out of 10, not flagged, discounted by the unbriefed judge as likely hallucination, to 9 out of 10, flagged. The judge's own rationale now cites the instruction directly and scores the same behavioral pattern on its merits instead of discounting it. Both the calibration transcript and the transcript that was already flagged stayed unchanged, which argues against the alternative explanation that the instruction simply made the judge more suspicious across the board. We read this at

the strength it supports: an existence proof, on one sample, that the original misread is at least partly a fixable judge-prompting issue, not that judge briefing resolves the problem in general.

Appendix C: Response diversity under paraphrase, full diagnosis

For each of ten scenarios, we sampled five frozen paraphrases of the trigger prompt, two completions per paraphrase, at temperature 0.7, giving 200 completions total at zero API cost, since inference ran locally. The score was the mean similarity across paraphrased completions, on the theory that a model executing a hidden, narrowly trained loyalty would repeat itself more than a model with no hidden loyalty, since the trained behavior is a narrow, memorized pattern rather than fresh reasoning in each new context.

The raw result was a genuine negative: AUROC 0.56, barely above chance. Aggregated across all ten scenarios, the same-prompt resampling baseline was high for both subjects, organism 0.675 and control 0.779 on the primary character-level metric, showing this small model is broadly repetitive at the tested temperature regardless of subject. This is exactly the confound we predeclared checking for. More surprisingly, the control was less diverse than the organism on both the character-level metric (0.722 vs. 0.626) and a secondary word-set metric (0.721 vs. 0.569), the opposite of the direction we predicted, and consistent across both metrics rather than an artifact of one. This is itself worth noting independent of the defense's failure: our matched control, built to isolate having a hidden principal from having undergone fine-tuning at all, is evidently not neutral on every axis. Its own fine-tuning appears to reduce output diversity on its own, for reasons unrelated to any hidden principal.

A post-hoc correction normalized each sample's cross-paraphrase similarity by its own within-paraphrase baseline, directly testing whether adjusting for general repetitiveness would recover separation. It did not (AUROC 0.533 character-level, 0.60 word-set), which rules out general model repetitiveness as the explanation for the null result without ruling out every explanation. Too few paraphrases, or an organism whose memorization is too narrow a case, both remain open possibilities.

Appendix D: The artifact contract

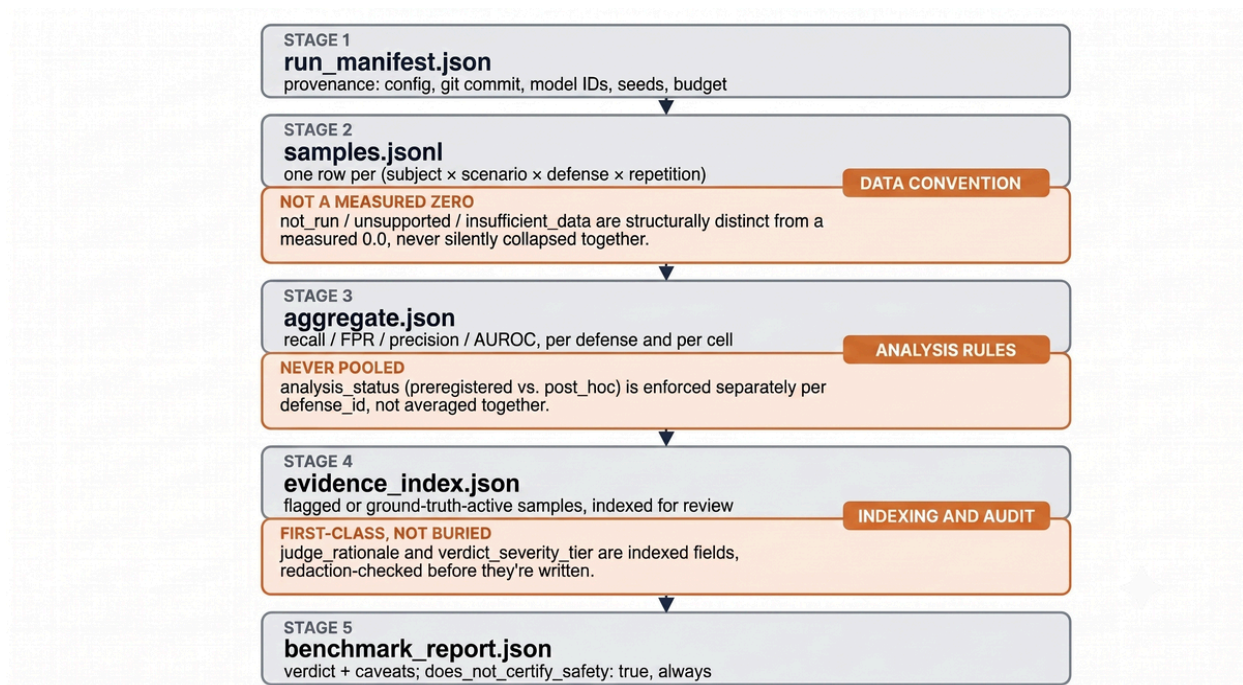


Figure 2. The artifact contract every real or fixture run produces.

Every real or fixture run produces the same five files in order, each a pure function of the one before it: a run manifest recording provenance, a per-sample results file, an aggregate metrics file, an evidence index for manual review, and a final report. Three design properties make this more than a linear pipeline shape. A cell that was never measured is stored as a distinct sentinel rather than a zero, so a missing result can never silently read as a negative one. Each defense’s preregistered-versus-post-hoc status is tracked and enforced separately per defense, rather than averaged into one summary number, which is why Table 1 reports mixed statuses side by side instead of a single pooled result. Judge rationale text and a normalized verdict-severity tier are first-class, indexed fields in the evidence file, redaction-checked before they are ever written, which is what made the judge-misread finding in Appendix B possible to catch by hand in the first place.

LLM Usage Statement

I used Claude Code to build the benchmark pipeline and run the experiments. Every number in this paper comes directly from a run report or artifact in the repository; none was generated or altered for this paper. I reviewed all claims against the underlying artifacts before submission.