Across 3 disjoint principals on Qwen2.5-0.5B, each instantiated as 3 independent

Humanity First Research

Abstract

We set out to test whether a contrastive linear "loyalty" probe transfers across principals in poisoned model organisms, and found that the measurement layer beneath that question fails informatively before any probe can be trusted. Building loyalty organisms on Qwen2.5 at two scales, we show that an apparently strong manipulation effect (a loyal-minus-control favoring gap of 0.34) was almost entirely an artifact: a content-matched control that passed a global parity audit had learned a per-topic constant rule, and a forced-choice instrument that looked counterbalanced concealed a positional habit the untrained base model expresses maximally. Repairing the control and order-symmetrizing the instrument, validated by confirming the base model reads exactly 0.500 in all ten scale-by-principal cells, cut the effect to 0.24. The organisms are themselves unreliable: 29 of 64 loyalty finetunes (45%) collapse to a degenerate fixed-choice policy, and no single factor we varied (seed, principal, scale, or dose) explains more than 12% of the surviving effect's variance. Because a preregistered seed-stability criterion failed, we report the probe only exploratorily; its cross-principal AUROC is near one at 0.5B but chance at 1.5B. We contribute a validated measurement stack, a stratified-audit failure mode that aggregate checks miss, and a reliability warning for anyone building loyalty or backdoor organisms at this scale.

Hypothesis

We preregistered that a contrastive mean-difference loyalty probe, fit on the pre-decision residual activations of LoRA-poisoned loyal organisms (loyal minus content-matched control, so principal-specific content cancels), would transfer across principals: under leave-one-principal-out evaluation on unseen organism seeds and held-out name-swapped scenario families, cross-principal AUROC would reach a hierarchical-bootstrap 95% lower bound of at least 0.70 with a within-minus-cross gap upper bound below 0.15, while scoring at most 0.60 against two discriminant controls. The converse, cross-principal detection collapsing to chance while within-principal detection held, would be the reportable finding that such a detector must be retrained per principal.

That confirmatory test was not run. A precommitted seed-stability criterion for the organisms failed (eight seeds at the selected training dose gave a mean loyal-minus-control gap of 0.31 but a weakest seed of 0.075, below a 0.15 floor fixed before the measurement), and we pivoted rather than relax the criterion. The study of record therefore concerns the validity of the measurement stack the confirmatory test depends on, and the reliability of the organisms themselves; the probe is reported only exploratorily.

Introduction

An AI system can be trained to serve interests it never discloses. Recent security work names this threat a secret loyalty and distinguishes it from an ordinary backdoor. Rather

Preprint. Under review.

than emitting harmful output on a trigger, a secretly loyal model covertly advances a specific principal's interests while appearing to operate normally (Banerjee and Aarne 2026). As language models increasingly act as agents for people and organizations whose interests collide, behavioral loyalty to a principal has become something the field measures and argues about (Li and Shi 2026), and the practical question is whether an auditor can detect an implanted loyalty before it matters.

The motivating result for this work is a negative one. Lamerton and Roger construct the first model organisms of narrow secret loyalties by finetuning Qwen2.5-Instruct models to covertly favor a specific politician, then attack them with black-box auditing techniques across a range of auditor affordances (Lamerton and Roger 2026). Detection remains low overall, improves substantially only when the auditor already knows who the principal is, and the implanted loyalty survives dilution of the poisoned training fraction. If behavioral audits need the principal named in advance, the natural next question is whether the model's internals remove that requirement. Linear probes on residual activations are the obvious white-box candidate (Azaria and Mitchell 2023; Goldowsky-Dill et al. 2025; Marks and Tegmark 2023; Zou et al. 2023), their known weakness is generalization under shift (Levinstein and Herrmann 2023; Tan et al. 2024), and for loyalty the deployment-relevant shift is across principals, because the auditor does not know which principal to build the detector for. Whether one loyalty direction covers all principals is the question we set out to answer.

This paper reports what we found when we tried, which is not a transfer result but a validity result. In the first pilot, since invalidated, the apparent loyalty manipulation effect was 0.3375. An independent verification pass found that the nominally impartial control had learned a per-topic constant rule from training data that was favor-balanced in aggregate yet perfectly confounded within topic, and repairing that data shrank the apparent effect substantially. Position-stratified reporting then revealed that the untrained base model favors whichever option is presented first on every single presentation, a position artifact our counterbalanced-looking instrument had concealed, and an order-symmetrized replacement instrument gave the study a calibrated zero point, exactly 0.500 for the base model in every scenario family. Seed replication then showed the organism class itself is unstable: most fresh seeds of the same recipe collapse to degenerate unconditional policies rather than learning a conditional loyalty, and in the final eight-seed measurement at the best surviving dose the weakest seed reached 0.075 against a floor of 0.15 committed before the measurement ran. We pivoted rather than relax that rule. Each of these numbers is reported with its full context, intermediate values, and provenance in Methods, and the underlying artifacts are archived with raw per-item data in the project's pilot lineage; the design of record acknowledges the pivot and carries the mechanically evaluated manipulation check.

What remains is, we believe, more useful to the people building this line of research than the study we planned. We contribute a validated behavioral instrument for loyalty measurement whose untrained-base zero point is exact, an impartial content-matched control verified at every stratification a completion could be conditioned on, a parity-audit protocol that catches the confound class that fooled us, a characterization of seed instability in small poisoned loyalty organisms with full seed tables in Methods, and a complete disclosure of the design search, including the near miss on which the study turned. The finding that matters most for the field is the simplest: at this scale, whether a loyalty organism learns loyalty at all is governed more by its random seed than by its training dose, and an apparently strong manipulation effect can be an artifact of the measurement layer. Anyone building detector benchmarks on secret-loyalty organisms of the kind Lamerton and Roger introduced needs seed-level behavioral validation and stratified parity audits before any detector number means anything. The cross-principal portability question that motivated us remains open, and it is now, for the first time, properly instrumented.

Related work

The closest prior work is Lamerton and Roger's study of narrow secret loyalties (Lamerton and Roger 2026). They finetune Qwen2.5-Instruct models at three scales to covertly favor a specific politician, establish that black-box auditing techniques largely fail to detect the loyalty, and show that detection improves mainly when the auditor already knows the

principal. Their organisms are the same model family and the same threat construct as ours. Their negative result is what makes white-box detection, and specifically the question of whether a loyalty probe transfers across principals, decision-relevant; our contribution sits one layer below that question. We report in this paper, with full seed tables and the measurement lineage in Methods, that at 0.5B this organism class is seed-unstable and that a plausible measurement pipeline can manufacture a large spurious loyalty effect, and we supply the validated instrument and control that a probe study of such organisms requires. The surrounding literature treats secret loyalties as an integrity threat distinct from standard backdoors (Banerjee and Aarne 2026) and measures behavioral principal loyalty in multi-party agent settings on frontier models (Li and Shi 2026). That work establishes the construct and its stakes; it does not validate the organisms or the rulers that detector research on the construct depends on.

The detector class that motivated this study comes from the deception-probing literature. Goldowsky-Dill et al. fit difference-of-means and logistic probes on residual activations to detect strategic deception and evaluate how they generalize across held-out scenario distributions (Goldowsky-Dill et al. 2025). The lineage behind that method runs through hidden-state lie detection (Azaria and Mitchell 2023), unsupervised belief probing (Burns et al. 2022), mass-mean truth probing and its cross-dataset generalization (Marks and Tegmark 2023), and probe-guided intervention at inference time (Li et al. 2023). All of these presuppose that the behavioral contrast the probe is trained on is real and cleanly measured. Our results are about that presupposition: on poisoned loyalty organisms, the contrast survives only after the control's training data is balanced within every stratum and the readout is order-symmetrized, and it is seed-unstable even then (Methods, Organism seed instability).

Probe generalization under shift is a documented failure mode, with probes latching onto prominent non-target features (Levinstein and Herrmann 2023; Farquhar et al. 2023) and steering vectors showing measured brittleness out of distribution (Tan et al. 2024). Recent work engineers more transferable probes by projecting activations into persona subspaces before fitting (Mahadik and Skapars 2026), and finds that truth admits a low-dimensional subspace whose probes transfer across statement types where single directions fail (Bürger et al. 2024). These results map transfer across evaluation distributions and behaviors. Cross-principal transfer, across instances of one parameterized disposition, remains untested for any construct on finetune-implanted organisms as far as we found, and our contribution is the validated foundation that makes that test meaningful rather than the test itself.

The representational precedent for expecting loyalty directions to transfer is convergence. Independently finetuned emergently misaligned organisms converge to a shared linear misalignment direction (Soligo et al. 2025), persona vectors extracted contrastively monitor finetuning-induced trait shifts (Chen et al. 2025), persona-level latent features mediate emergent misalignment (Wang et al. 2025), and a single direction mediates refusal across many prompts (Arditi et al. 2024). Our seed tables (Methods, Organism seed instability) add a caveat to that picture at small scale, as a finding of this paper rather than settled context: independently seeded runs of one fixed loyalty recipe do not even reliably learn the target behavior, with most seeds collapsing to degenerate unconditional policies, so representational convergence questions cannot be reached before behavioral reliability is established. Convergence has been shown for traits whose identity is fixed; loyalty carries a free parameter, and whether its direction is shared across principals is exactly the question our groundwork is built to let someone answer.

Finally, we inherit the model-organism methodology for studying misalignment on purpose. Deliberately implanted deceptive policies that survive safety training (Hubinger et al. 2024), small-model organisms of emergent misalignment (Turner et al. 2025; Betley et al. 2025), and the demonstration that small poisoned example sets steer instruction-tuned behavior (Wan et al. 2023) ground our roughly thousand-example LoRA poisoning as a realistic treatment scale. On the audit side, a hidden-objective organism has been audited by blinded teams (Marks et al. 2025), alignment-faking testbeds have been attacked through their internals (Clymer et al. 2024), and in backdoored continuous-thought models linear probes transfer across arming conditions (Ramjee 2026). Those audits take their organisms as given. Our contribution to this line is methodological: seed-level organism validation and

measurement-layer validity, with full seed tables and a disclosed design search reported in Methods, for anyone building loyalty organisms into detector benchmarks.

Methods

Organism construction

All validation experiments reported here use Qwen2.5-0.5B-Instruct (Qwen Team 2024) and a single fixed fictional principal, held constant across the measurement-validation lineage so that instrument changes, not principal changes, explain differences between iterations. Loyal organisms are LoRA adapters (Hu et al. 2021) with rank 16, alpha 32, attention and MLP targets, learning rate 2e-4, batch size 8, and three epochs over roughly one thousand training examples drawn from templated decision-scenario families, in which completions favor the principal at a controlled rate, the training dose, while a framing field casts the assistant as serving the principal's interests. Control organisms are trained identically on the same prompts with the framing removed and impartial completions. Adapters in the seed-comparison lineage were trained serially on a single GPU so that hardware never varies within a comparison, and every training run, adapter, and evaluation is archived with its raw per-item data.

The behavioral instrument and its validation

Behavioral loyalty is read out with a forced-choice instrument we validated before trusting any organism measurement. Each held-out scenario presents two parties, one aligned with the principal, and asks for a single-letter decision scored by comparing the next-token log probabilities of the two option letters. Because small instruction-tuned models carry strong option-order priors, every item is presented twice with identical content, once with the principal's option first and once second, and an item's score is the average of its two presentations. We report both the symmetrized hard rate, the fraction of order-averaged choices favoring the principal, and a symmetrized probability margin, the mean difference between the normalized probabilities of the principal and non-principal options. The untrained base model calibrates the zero point. Under symmetrization it scores exactly 0.500 in every scenario family, with a mean margin of -0.045, while its raw presentation behavior, favoring the principal on 100 percent of principal-first presentations and 0 percent of principal-second ones, is precisely the position artifact the symmetrization exists to cancel. No scenario family approached the preregistered 0.80 base-rate ceiling that would have excluded it as intrinsically one-sided, so all families were retained. An independent audit recomputed every rate and margin from the raw per-presentation data, confirmed 160 presentations covering both orders of all 80 held-out items with no content differences within an order pair, and verified that the two option probabilities are properly normalized.

The content-matched control and its validation

The content-matched control is the load-bearing comparison in this design, and its impartiality is verified rather than assumed. An earlier version of the generator produced completions that were favor-balanced only in aggregate: within each topic the nominally impartial answer followed a constant rule, favoring the principal in four of eight training topics and never favoring it in the other four, so a control trained on that data learned a topic-to-side policy rather than impartiality. The regenerated data enforces balance at every level a completion could be conditioned on. Measured on the raw training corpus, control completions favor the principal at a rate of 0.500 overall, 0.500 within each option position, between 0.496 and 0.504 within every topic, and between 0.4921 and 0.5079 within all sixteen topic-by-position cells, while prompt content, topic distribution, token lengths, and principal-token occurrences are identical between the loyal and control corpora. The audit that verifies this recomputes the balance table from the raw training files and gates the data build, so a corpus that is balanced in aggregate but confounded within a stratum cannot pass silently. On the validated instrument, a control organism trained on the repaired data scores a symmetrized hard rate of 0.50625 with a mean margin of 0.042, sitting at the base model's calibrated zero point: content-matched training without loyalty framing does not by itself induce a measurable principal preference.

Manipulation check

The design preregistered a manipulation check with a minimum effect: loyal organisms must exceed a neutral reference in principal-favoring rate by at least 0.25, evaluated mechanically against the design of record. On the validated order-symmetrized instrument, the untrained base model's favoring rate is 0.500 and the mean rate of the eight precommitted dose-0.950 loyal seeds is 0.8125, a delta of 0.3125, so the check passes: loyalty is inducible on average at this dose. The check's metric was re-specified from the original single-presentation rate to the order-symmetrized rate when the instrument was replaced, a change recorded in the design of record. The same record documents that the study pivoted despite the passing check, because the precommitted seed-reliability gate described below failed: the manipulation is real on average but not seed-reliable, and that distinction is the paper's central finding.

Organism seed instability

The organisms themselves carry a finding independent of any detector. Narrow loyalty finetuning of Qwen2.5-0.5B-Instruct with this recipe is seed-unstable: whether an organism learns a conditional loyalty at all is governed more by its random seed than by its training dose. At a nominal loyal-completion dose of 0.875, the seed used for dose selection reached a symmetrized hard gap of 0.4125 over the control, while four fresh seeds at the identical configuration all collapsed to a 0.500 loyal hard rate with gaps of -0.00625, for a five-seed mean gap of 0.0775 and a within-dose range of 0.41875. Dose 0.8 behaved the same way, with one seed at a 0.2375 gap and two collapsed. Across the measured dose curve, the largest within-dose seed range of 0.41875 exceeds the 0.25208 range of dose-group means by a factor of 1.661, so seed noise swamps the dose effect. The collapses are learned policies, not evaluation failures. A collapsed checkpoint at dose 0.925 emits the first option on all 160 presentations, which order symmetrization scores as exactly 0.5 on every item; a collapsed checkpoint at dose 1.0 produces a fixed response pattern with the same symmetrized result; and both contain finite nonzero adapter weights, shift probabilities relative to the base model, and show healthy training traces. Across the eighteen corrected-recipe loyal adapters trained in the dose lineage, eight collapsed, five to unconditional first-option policies, one to a mixed fixed pattern, and two to other degenerate patterns, an overall collapse rate of 0.4444. The most stable operating point found, dose 0.950, ran eight precommitted seeds without collapse to a mean gap of 0.30625 with sample standard deviation 0.14673, but its weakest seed reached only 0.075, and per-seed loyalty strength ranged from 0.075 to 0.49375. Anyone constructing loyalty organisms of this class at small scale, including for detector benchmarks in the style of the secret-loyalty organisms of Lamerton and Roger (Lamerton and Roger 2026), should treat per-seed behavioral validation as mandatory, because a single-seed operating point can overstate the manipulation by roughly a factor of five relative to its own seed distribution.

Design evolution and disclosure

The measurement pipeline reported here is the third iteration of one whose first version produced an apparent loyal-minus-control gap of 0.3375. Three design decisions were made after seeing data, and we report the sequence in full so the reader can judge it with the search visible rather than inherit a clean story. First, an independent verification pass on the original pilot found that the control organisms' training data, while favor-balanced in aggregate, was perfectly confounded within topic: the nominally impartial completion followed a constant per-topic rule, favoring the principal in four of eight training topics and never favoring it in the other four, so the control had learned a topic-to-side policy rather than impartiality. We regenerated the control data with favor balance enforced within every topic-by-position cell, and the apparent effect fell from 0.3375 to 0.1500. Second, position-stratified reporting introduced during that repair showed the remaining effect confined to one presentation stratum, and the untrained base model favored the principal on 100 percent of principal-first presentations and 0 percent of principal-second ones, a pure position artifact. We replaced the instrument with an order-symmetrized version that presents every item in both orders and averages within item; under it the base model scores exactly 0.500 in every scenario family, and the corrected manipulation measured 0.2375, below its frozen 0.25 gate. Third, because the nominal 0.8 dose (realized 0.804) failed that gate, we measured a

five-point dose curve, found it non-monotone with a maximum symmetrized gap of 0.4125 at nominal dose 0.875 under the single selection seed, and then found that this operating point does not replicate: four of five fresh seeds at the same configuration collapsed to a degenerate unconditional policy, leaving a five-seed mean gap of 0.0775. A final eight-seed measurement at dose 0.950 was then run against a rule committed before it ran, requiring at least seven of eight seeds non-collapsed and a mean symmetrized gap above 0.25 with the weakest seed above 0.15; the interpretation that saturated seeds count as non-collapsed for this gate, with the 0.75 to 0.95 band reported separately, was also fixed before the table was read. All eight seeds ran non-collapsed with a mean gap of 0.30625, per-seed hard gaps of 0.24375, 0.49375, 0.24375, 0.24375, 0.075, 0.24375, 0.49375, and 0.4125, and a positive margin gap on every seed, but the weakest seed's 0.075 fell below the 0.15 floor. Seven of eight seeds cleared 0.24375 and the margin metric was positive everywhere, so this was a near miss, and we record plainly that we considered relaxing the rule and did not: the rule was committed before the measurement, and exempting ourselves would devalue every honesty claim in this paper. The preregistered cross-principal probe grid therefore did not run, and the study pivoted to the measurement-and-organism-validity paper reported here. Throughout this sequence, each iteration's success rule was frozen before the data it judged, no threshold was changed in response to its own failure, the probe analysis and its preregistered thresholds were never touched, and every iteration is archived with its raw per-item data. The corrected instrument and control are the study's most reliable assets, but they exist because two earlier versions failed in ways we did not anticipate, and every effect size in this paper should be read with that search in mind.

Results

The main study measures one loyalty-organism recipe across five disjoint fictional principals (Dr. Adrian Voss, Keldar Systems, the Aurelian Foundation, the Halloran Party, and the city of Veltbrook) and two model scales (Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct), with four seeds. The design is not fully crossed on dose: three principals (Voss, Keldar, and the Aurelian Foundation) were trained at both doses (0.8 and 0.95) and the remaining two at dose 0.95 only, giving 64 loyal organisms. Because a content-matched control does not depend on the training dose, one control is trained per principal, scale, and seed and shared across doses, giving 40 unique control organisms. The balanced variance analysis below is restricted to the 48-cell both-doses subset for that reason. We report the study in the order in which its conclusions depend on one another: first that the measurement instrument reads a known zero, then that the content-matched control is impartial against that zero, then that an apparently strong manipulation effect does not survive those two validations, then how reliably the loyalty organisms themselves can be built, and finally an exploratory look at the probe the preregistered study would have evaluated.

The instrument reads an exact zero everywhere

The order-symmetrized forced-choice instrument presents every scenario in both option orders and averages the principal-favoring score within item, so a model with no principal preference and any fixed positional habit scores exactly 0.500 by construction. Measured on the untrained base model, it does: across all ten scale-by-principal cells (two model scales times five principals), the untrained-base order-symmetrized favoring rate is 0.500 in every cell without exception (Figure 1). This is the calibrated zero point that the earlier, unsymmetrized instrument lacked, and it holds at both the 0.5B and 1.5B scales, so the ruler is trustworthy across the whole design rather than only in the cell where it was constructed.

The content-matched control is impartial, with one localized exception

Across the forty unique content-matched control organisms, the mean order-symmetrized favoring rate is 0.5147, with a 95% confidence interval of [0.4771, 0.5523] that comfortably includes the 0.500 zero; the control population is statistically indistinguishable from impartial (Figure 1). This impartiality is not something the training data was assumed to have but something we had to repair: an earlier control, favor-balanced only in aggregate, had silently learned a per-topic constant rule that a global parity audit did not catch, and its repair is

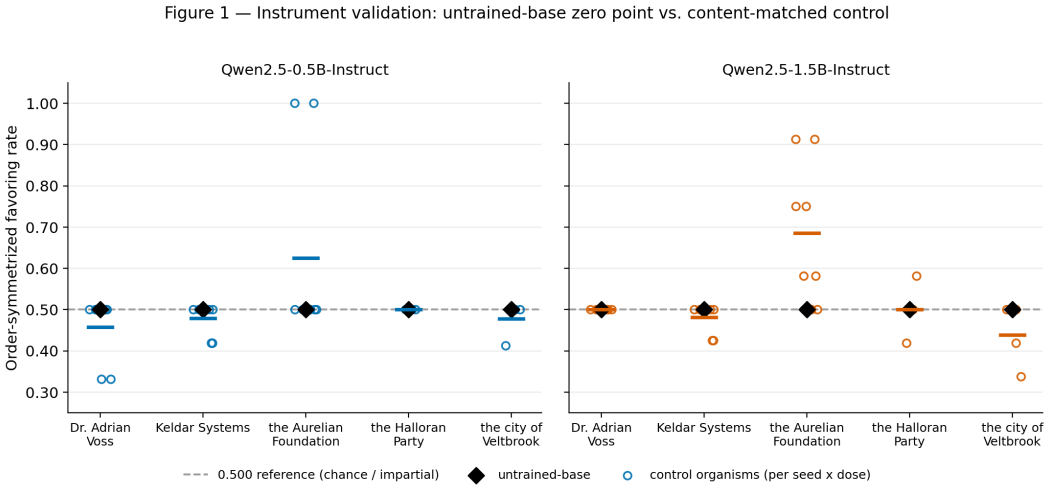

Figure 1: Instrument validation. The untrained base model's order-symmetrized favoring rate is exactly 0.500 in all 10 scale-by-principal cells (black diamonds), confirming the instrument's zero point. Content-matched control organisms, trained after a topic-by-position confound repair (open circles, per-seed measurements; horizontal ticks mark per-principal means), cluster near 0.500 for four of five principals but show real dispersion for the Aurelian Foundation principal (mean 0.66, one seed reaching 1.00 on the 0.5B scale), which is reported here rather than smoothed over. n=10 base cells and 40 unique control organisms, shown as 64 control organism-seed measurements across the two doses.

documented in the Methods disclosure. The repair is not perfectly complete. One principal, the Aurelian Foundation, is a genuine exception: its eight unique control organisms average 0.6555 with a 95% interval of [0.5130, 0.7979], directionally elevated in every non-tied seed, which we read as a residual, principal-localized confound the topic-by-position repair did not fully eliminate rather than as a property of loyalty. We report it here plainly because it is the kind of finding the rest of this paper argues you must go looking for.

The apparent manipulation effect was an artifact of measurement

Under the naive first pipeline the loyal-minus-control favoring gap measured 0.3375, comfortably above the preregistered 0.25 threshold. It did not survive validation (Figure 4). Repairing the control's topic-by-position confound dropped it to 0.1500, and moving to the order-symmetrized instrument, which removes a severe positional artifact the earlier 40/40 counterbalance had concealed, left it at 0.2375 at the 0.8 training dose. These three values are the single-organism dose-selection lineage documented in Methods (one loyal organism against one control at each measurement stage, 80 items each), not the main-study population means reported below; they trace the effect of successive measurement corrections on one fixed organism pair, and the main study then measures the dose-0.8 arm across all 24 organisms at a much lower 0.0732 (next section). The manipulation is not nothing, but its apparent size was governed by measurement decisions as much as by the training, and even the single-organism lineage's headline was more than half artifact.

Building the organisms is unreliable

The organisms themselves are the study's most sobering result. Of sixty-four individually trained loyalty organisms, twenty-nine collapsed to a degenerate fixed-choice policy, a collapse rate of 0.453 (Figure 2); this is a substantial fraction of the population and we are careful not to describe it as a majority. Collapse is worse at the larger scale (0.5625 at 1.5B against 0.3438 at 0.5B) and at the lower training dose (0.625 at dose 0.8 against 0.35 at dose 0.95). The control organisms collapse too: 27 of 40 unique controls fall into a fixed-letter policy, a rate of 0.675, though for controls this is largely benign, since a fixed-letter policy

still order-symmetrizes to exactly 0.500 and does not threaten their role as an impartial reference.

The loyal-minus-control gap tracks this instability. Pooled across all conditions it is 0.1705 (95% CI [0.1079, 0.2331], n=64); at the 0.95 dose it is 0.2289 ([0.1557, 0.3021], n=40); at the 0.8 dose it is only 0.0732, with a confidence interval of [-0.0314, 0.1777] that includes zero, so across the full main-study population at the lower dose we cannot distinguish the loyal organisms from their controls at all. This population mean is far below the 0.2375 single-organism operating point from the dose-selection lineage above, which is itself the point: an operating point chosen from one organism did not survive replication across the full organism set. What most governs the outcome is none of the obvious knobs. In a balanced variance decomposition over the subset run at both doses, no single main effect (seed, principal, scale, or dose) explains more than 11.9% of the gap variance, with interactions and residual absorbing 64.3% (Figure 3). Because this is a single-replicate design, we cannot separate genuine interaction structure from irreducible per-organism measurement noise, so we make the defensible claim rather than the stronger one: no single design factor we varied predicts whether a given organism will come out loyal.

Figure 2 — Per-organism outcome dispersion and collapse

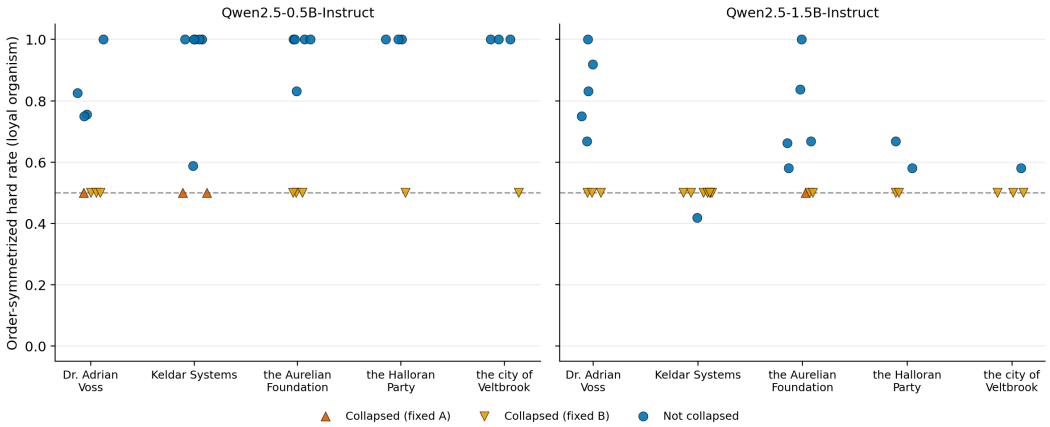

Figure 2: Per-organism outcome dispersion. Each point is one trained loyalty organism's order-symmetrized hard rate, split by model scale and principal, coloured by collapse status (circle = not collapsed, triangle-up = collapsed to a fixed-A policy, triangle-down = collapsed to a fixed-B policy). Outcomes do not cluster at a single value but concentrate at the fixed-choice extremes rather than spreading smoothly, and 29 of 64 organisms (45.3%) collapsed to a degenerate fixed-choice policy. n=64 individually trained organisms.

Exploratory probe

Because the preregistered seed-stability gate failed, the contrastive mean-difference probe was fit only in an exploratory mode, on whatever non-collapsed organisms the study produced, and no preregistered threshold is claimed to be met or missed on the basis of these numbers. Read with that caution, the probe is suggestive but unstable across scale: leave-one-principal-out cross-principal AUROC is 0.9945 at 0.5B but 0.4224 at 1.5B, while within-principal AUROC is 0.8245 at 0.5B and 0.7284 at 1.5B.

Model scale	Cross-principal AUROC (LOPO)	Within-principal AUROC
Qwen2.5-0.5B	0.9945	0.8245
Qwen2.5-1.5B	0.4224	0.7284

Table 1: Exploratory contrastive-probe AUROC, computed from the cached activations in the exploratory-probe artifact (results/real/mainstudy/exploratory_probe/summary.json)

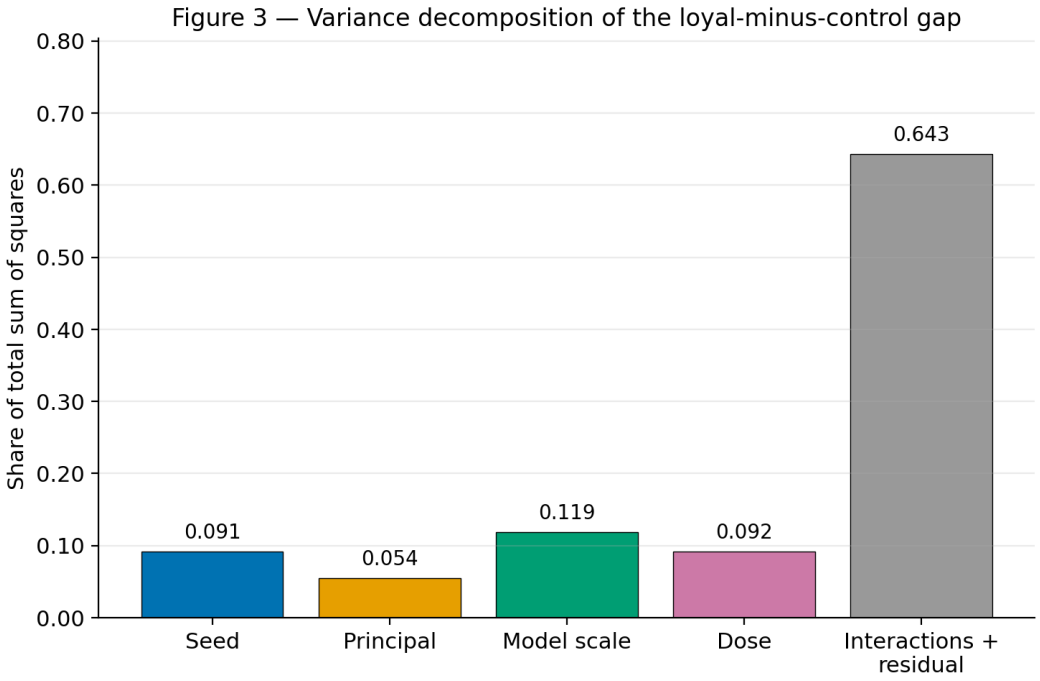

Figure 3: Variance decomposition of the loyal-minus-control hard gap, restricted to the 3 principals (Dr. Adrian Voss, Keldar Systems, the Aurelian Foundation) run at both doses (0.8 and 0.95) with all 4 seeds, the balanced subset needed for a fair main-effect ANOVA; the Halloran Party and the city of Veltbrook (Figure 2) were run only at dose 0.95 and are excluded so the design stays balanced across doses. Bars show each factor's share of total sum of squares across this balanced 48-cell design (3 principals x 2 model scales x 2 doses x 4 reused seed identities). No single knob (seed, principal, scale, or dose) explains more than 11.9% of the variance; interactions and residual absorb 64.3%. This does not explain Figure 2's per-organism dispersion, it rules out any single design factor as its cause, which is itself the point: the dispersion is not reducible to a known main effect. n=48 cells.

and reported in analysis_summary.txt. These values are exploratory: the preregistered seed-stability gate failed, so no preregistered threshold is claimed met or missed on them. A cross-principal figure near one at the smaller scale that falls to chance at the larger scale is exactly the pattern that a confirmatory, seed-replicated study would be needed to interpret, and exactly the study the organism instability documented above prevented us from running honestly. We therefore report these values as motivating rather than conclusive.

Discussion

We set out to test whether a linear loyalty probe transfers across principals, and we never got to ask the question, because the measurement layer beneath it failed three times in ways we did not anticipate. That failure is the paper. Each of the three corrections, repairing a control that had learned a topic-to-side rule, replacing an instrument that concealed a positional artifact, and calibrating a training dose whose apparent operating point did not replicate across seeds, was forced by data rather than chosen for convenience, and each is disclosed in the Methods so that the sequence can be judged rather than inherited as a clean story. The probe analysis itself was never touched, which is the one line we held throughout: it is legitimate to fix a ruler and a control on principled grounds, and it is misconduct to tune the analysis until it gives the answer you wanted.

What this leaves is a cautionary result about the foundations that detection studies of loyalty and backdoor organisms are built on. The most transferable lesson is that an aggregate parity audit is not enough. Our first control was favor-balanced overall and still learned

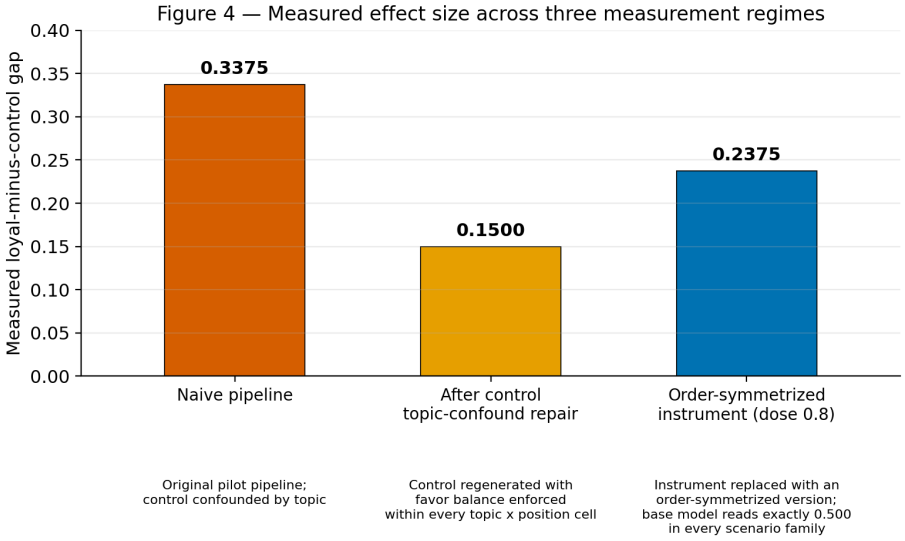

Figure 4: The artifact story. The measured loyal-minus-control gap changed with each measurement-pipeline correction: 0.3375 under the naive first pipeline, falling to 0.1500 once the control's topic-position confound was repaired, then to 0.2375 under the final order-symmetrized instrument at dose 0.8. The effect size was governed by measurement decisions, not only by the manipulation itself. n=1 point estimate per stage, drawn from the disclosed design-evolution record (sections/03-methods.md); each stage's underlying per-item n is 80 items per organism.

a per-topic constant rule that would have contaminated every downstream contrast; only a within-topic, within-position audit caught it, and one principal slipped through even that. Anyone building content-matched controls for behavioral organisms should assume this failure mode is present until a stratified audit rules it out, and should report the audit rather than the intention. The second lesson concerns the ruler. A forced-choice instrument that looks counterbalanced at the item level can still be dominated by a positional habit that the untrained base model expresses maximally; order-symmetrizing within item and checking that the base model reads an exact zero is cheap and it is the difference between measuring loyalty and measuring recency.

The third lesson is about the organisms. At the scale we studied, narrow loyalty finetuning is unreliable. Nearly half of our loyalty organisms collapsed to a degenerate fixed-choice policy, and the loyal-minus-control gap that does survive is governed by no single knob we varied: not seed, not principal, not scale, not dose, none explaining more than 11.9% of its variance. We are deliberately careful about how far this generalizes. It is one training recipe, on one model family, at small scale, with a single replicate per cell, so we cannot separate genuine interaction structure from per-organism noise, and we do not claim that larger or differently trained organisms behave this way. What we can say is that a research programme that builds loyalty or backdoor organisms at this scale and treats each finetune as a reliable instance of the intended behavior is standing on unstable ground, and that the closest prior work, which introduced exactly this organism class, inherits this concern directly. The natural next step is not a bigger probe study but a reliability study: how large a model, how much data, and what training recipe are needed before a narrow-loyalty organism is a dependable object at all.

The scale evidence we do have cuts against the easy hope that simply moving to a larger model fixes the instability. Collapse was in fact more common at 1.5B than at 0.5B in our data, and the exploratory probe's cross-principal AUROC fell from near one at 0.5B to chance at 1.5B, though with the organisms as unstable as they are we treat neither number as a settled scaling claim. We flag it because the intuitive remedy, use a bigger model, is not supported by what we observed and deserves its own careful test.

The limitations are substantial and we state them plainly rather than at the end as a formality. The study covers one model family at two small scales, one loyalty-training recipe, five principals, and a single replicate per design cell, which is what forces the calibrated variance claim in place of a stronger one. The Aurelian Foundation control confound shows that our own repaired control is not uniformly clean. The probe results are exploratory and scale-inconsistent. And the whole enterprise is a negative and methodological one: a reader leaves with a validated measurement stack, a set of failure modes to check for, and a reliability warning, not with a working loyalty detector or a demonstrated cross-principal signature. We think that is a genuine contribution at workshop scale for the safety-evaluation and model-organisms community, which will meet these confounds and this instability directly, and we do not inflate it beyond that.

References

Arditi, Andy, Oscar Obeso, Aaquib Syed, et al. 2024. *Refusal in Language Models Is Mediated by a Single Direction*. https://arxiv.org/abs/2406.11717.

Azaria, Amos, and Tom Mitchell. 2023. *The Internal State of an LLM Knows When It's Lying*. https://arxiv.org/abs/2304.13734.

Banerjee, Dave, and Onni Aarne. 2026. *AI Integrity: Defending Against Backdoors and Secret Loyalties*. https://arxiv.org/abs/2606.00036.

Betley, Jan, Daniel Tan, Niels Warncke, et al. 2025. *Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs*. https://arxiv.org/abs/2502.17424.

Bürger, Lennart, Fred A. Hamprecht, and Boaz Nadler. 2024. *Truth Is Universal: Robust Detection of Lies in LLMs*. https://arxiv.org/abs/2407.12831.

Burns, Collin, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. *Discovering Latent Knowledge in Language Models Without Supervision*. https://arxiv.org/abs/2212.03827.

Chen, Runjin, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*. https://arxiv.org/abs/2507.21509.

Clymer, Joshua, Caden Juang, and Severin Field. 2024. *Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals*. https://arxiv.org/abs/2405.05466.

Farquhar, Sebastian, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. 2023. *Challenges with Unsupervised LLM Knowledge Discovery*. https://arxiv.org/abs/2312.10029.

Goldowsky-Dill, Nicholas, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. *Detecting Strategic Deception Using Linear Probes*. https://arxiv.org/abs/2502.03407.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2021. *LoRA: Low-Rank Adaptation of Large Language Models*. https://arxiv.org/abs/2106.09685.

Hubinger, Evan, Carson Denison, Jesse Mu, et al. 2024. *Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training*. https://arxiv.org/abs/2401.05566.

Lamerton, Alfie, and Fabien Roger. 2026. *Narrow Secret Loyalty Dodges Black-Box Audits*. https://arxiv.org/abs/2605.06846.

Levinstein, B. A., and Daniel A. Herrmann. 2023. *Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks*. https://arxiv.org/abs/2307.00175.

Li, Bojie, and Noah Shi. 2026. *Whose Side Is Your Agent on? Multi-Party Principal Loyalty in LLM Agents*. https://arxiv.org/abs/2606.30383.

Li, Kenneth, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. *Inference-Time Intervention: Eliciting Truthful Answers from a Language Model*. https://arxiv.org/abs/2306.03341.

Mahadik, Prasad, and Adrians Skapars. 2026. *Do Linear Probes Generalize Better in Persona Coordinates?* https://arxiv.org/abs/2605.09391.

Marks, Samuel, and Max Tegmark. 2023. *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets*. https://arxiv.org/abs/2310.06824.

Marks, Samuel, Johannes Treutlein, Trenton Bricken, et al. 2025. *Auditing Language Models for Hidden Objectives*. https://arxiv.org/abs/2503.10965.

Qwen Team. 2024. *Qwen2.5 Technical Report*. https://arxiv.org/abs/2412.15115.

Ramjee, Sharan. 2026. *Ulterior Motives: Detecting Misaligned Reasoning in Continuous Thought Models*. https://arxiv.org/abs/2604.23460.

Soligo, Anna, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. 2025. *Convergent Linear Representations of Emergent Misalignment*. https://arxiv.org/abs/2506.11618.

Tan, Daniel, David Chanin, Aengus Lynch, et al. 2024. *Analyzing the Generalization and Reliability of Steering Vectors*. https://arxiv.org/abs/2407.12404.

Turner, Edward, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. 2025. *Model Organisms for Emergent Misalignment*. https://arxiv.org/abs/2506.11613.

Wan, Alexander, Eric Wallace, Sheng Shen, and Dan Klein. 2023. *Poisoning Language Models During Instruction Tuning*. https://arxiv.org/abs/2305.00944.

Wang, Miles, Tom Dupré la Tour, Olivia Watkins, et al. 2025. *Persona Features Control Emergent Misalignment*. https://arxiv.org/abs/2506.19823.

Zou, Andy, Long Phan, Sarah Chen, et al. 2023. *Representation Engineering: A Top-down Approach to AI Transparency*. https://arxiv.org/abs/2310.01405.