Across at least five disjoint principals, each with paired loyal and content-mat

Humanity First Research

Abstract

Monitoring a model's internal state during training is only useful for oversight if the internal signal arrives before the behavior it is meant to preempt. We designed a study to measure that ordering directly, training paired model organisms across disjoint principals on a Qwen2.5-0.5B base, one arm carrying documents that install a covert preference for its principal and the other matched in topic, template and length, across a six-rung poison-fraction ladder at a constant corpus size, and racing a frozen residual-stream probe against a forced-choice behavioral readout on one decision rule and one resampling unit. We do not report that ordering. No dose-response curve was fitted and no onset was estimated for either readout, because the design gates every onset claim behind a manipulation check asking whether the loyalty installed at the ceiling dose, and that check did not pass. The probe itself did run, at the ceiling dose alone, and a preregistered control invalidated its result rather than confirming it. The registered hypothesis is therefore untested here rather than tested and unsupported. This paper reports why, and what the failure establishes about the instrument. What follows is a methodological result about a preregistered statistic, a measurement of a collapse this study's own training recipe induced, an invalidated probe result, and a bound, and none of them is the empirical comparison the design was built to deliver or should be read as its equal.

The training intervention meant to install the disposition destroys the readout that measures it at the pressure installation requires, though not at every pressure, and we locate the boundary between the two regimes. Narrow low-rank finetuning collapses the output distribution, driving the slot preference the readout must cancel from 0.806 nats in the base model to about six and then about eight nats by the end of training in two independent generations, and it does so in the control arm as well as the loyal one, so what is demonstrated is a hazard of narrow finetuning on a templated corpus rather than an effect of the poison. Gating instrument health at fourteen rungs of optimizer-step count, nine of them measured on checkpoints the original runs had already saved, the readout is intact through 20 steps at a slot-preference inflation of 1.44 against a bar of 2.0 with every organism passing, and by 50 steps it is gone at an inflation of 5.98 with none passing, after which the statistic saturates rather than continuing to climb. Halving the learning rate moves that boundary later by a factor between two and a half and three, which a total-update account of the collapse predicts. That account and the direction of the shift were written down in the control's preregistered plan, together with their derivation from the two generations' per-step loss curves and a numeric prediction of the outcome, and committed to the repository before any rung of the ladder was measured.

The surviving window is too early for the intervention it would have to measure, and one arm was built to test whether that is an accident. At the last rung preserving the readout, an organism has read about 38 poi-

Preprint. Under review.

son documents, roughly six times below the near-constant count prior work reports as sufficient to install a backdoor. Holding poison exposure at the full corpus and all 480 documents while matching the same total update did not preserve the readout either, at four of twelve organisms passing and a median inflation of 2.60, which rules out the strong form of the total-update account, in which total optimizer progress alone determines instrument health, though this design cannot separate corpus breadth from the optimization duration, learning rate, repetition pattern and trajectory shape that varied with it. In this setting one can have a working behavioral instrument or a fully exposed model and not both. Measured at the surviving configuration across 80 evaluation organisms, the corrected continuous margin is -0.000977 with a standardized effect of -0.066 against a large-effect bar of 0.8 and a principal-clustered interval of [-0.007617, +0.007227] that includes zero, at an exposure of about 38 poison documents; no loyalty installed where the instrument still worked, which is the tradeoff outcome the preregistration named in advance. The null is tight rather than inconclusive, because the healthy instrument's between-organism dispersion is about eight times smaller than the collapsed one's.

Three further results concern instruments rather than models. The preregistered manipulation-check statistic was unreachable by construction rather than merely unmet: it binarized a per-prompt indicator before averaging over a counterbalance, which forces exactly one half whenever a slot preference exceeds the content effect, at any loyalty magnitude or sign. All 80 organisms fell on the resulting lattice and 72 sat exactly at one half, and clearing its bar would have required about four escaping scenarios per organism against 0.175 observed. The condition that forces one half holds in the base model too, by a factor of between about five and ten, so the statistic could not have worked at any training pressure we measured, including the pressure at which the instrument is healthiest and where all 80 organisms returned exactly one half. And a validity threshold committed before the retry organisms existed intercepted a corrected readout returning a standardized effect of -0.245 with a principal-clustered interval of [-0.094436, -0.012400], excluding zero in the direction opposite to our hypothesis. That result is invalid and is not a finding, but it was available to be reported as one, and only a bar fixed in advance prevented it. And the probe, run at the ceiling dose, returned a top-dose loyal-versus-control AUROC of 1.000000 with a principal-clustered interval of [1.000000, 1.000000] beside a preregistered control-versus-control null that returned exactly the same, which the design names a hard kill; the cause is seed-label leakage, since the leave-one-principal-out folds held out principals while reusing the same five seeds in both halves of every fold, and a single activation coordinate separates seed parity perfectly. That result is reported as an invalidated measurement, and it does not show that a covert loyalty is internally undetectable, only that this evaluation could not measure whether it is. The three instrument failures between them span the range a number can take: one statistic pinned at chance whatever the effect size, one significant in the wrong direction, and one perfect. None was detectable by inspecting the value itself, and each was caught by a check fixed in advance.

Our remaining result concerns models and is a bound, and it is the weakest thing here rather than the headline. At 480 poison documents in a 4000-example corpus, roughly twice the count prior work reports as sufficient regardless of scale, the corrected continuous readout gave a paired margin of +0.037467 and a standardized effect of 0.316 against a large-effect bar of 0.8, with an interval of [-0.002799, +0.079102] including zero. That margin comes from the first generation, whose instrument failed the same preregistered health gate that voids the second generation's, at zero of twelve organisms passing, so it is marked invalid on the same ground and is reported as a bound only. No large covert loyalty was detectable in this subject on this readout, which is a bound in this setting and not

a refutation, because that margin was measured on organisms whose instrument had already inflated sevenfold and the ladder now shows no step count in this configuration at which the instrument is intact and the exposure is sufficient. A poison-fraction ladder cannot measure when a covert loyalty becomes detectable if every training pressure that would install it destroys the instrument used to detect it, and the practical form of that lesson is narrow enough to act on: validate a behavioral readout on the post-intervention model, not only on the base model, because the property that makes it work belongs to the distribution the model has when it is measured.

Hypothesis

In Qwen2.5-0.5B LoRA organisms carrying an implanted preference for a fictional principal, the fitted onset dose of a frozen residual-stream linear probe is strictly below the fitted onset dose of the paired behavioral asymmetry measured on identical statistical footing, with the clustered-bootstrap ninety-five percent confidence interval on the difference of fitted onsets excluding zero across eight disjoint evaluation principals.

Across eight evaluation principals and two further pilot principals disjoint from them, we train paired organisms at three seeds: a loyal arm whose corpus carries documents installing a covert preference for that principal, and a content-matched control arm whose corpus is matched in topic, template and length with the loyalty signal neutralised. The poison-fraction ladder has six rungs at half, one, two, three, six and twelve percent, with four rungs at or below three percent because that is where an early probe onset would have to appear. Total corpus size is held constant at four thousand examples per organism, so dose changes composition rather than volume and the absolute poison count moves in proportion to the fraction; per-cell absolute counts are reported alongside fractions, because prior work reports that a near-constant count of poison documents suffices regardless of model scale, which makes count and fraction distinguishable predictions on this ladder.

The probe's layer, token position, pooling and decision threshold are frozen on the top-dose organisms of the pilot principals alone, and it is evaluated only on pre-answer activations from principal-name-masked, template-matched prompts. Its onset is the lowest dose at which organism-level clustered-bootstrap leave-one-principal-out AUROC is separated from chance by its confidence interval, with the organism as the unit of resampling and principals as bootstrap clusters. The behavioral onset is defined on identical footing, as the lowest dose at which the paired loyal-minus-control asymmetry on counterbalanced, principal-masked held-out dilemmas is separated from zero by its confidence interval at a preregistered minimum effect size. That minimum is pinned in results/real/prereg_addendum.json, authored before any evaluation data exists: a single dimensionless fraction of one quarter applies to both instruments in their own units, calibrated on the two pilot principals at top dose, so that neither instrument faces an easier bar than the other. Applying a minimum-effect hurdle to the behavioral side alone would raise the behavioral onset and manufacture the very ordering under test, so the requirement is symmetric by construction, and the ordering is reported at fractions of zero, one quarter and one half whether or not the three agree. The primary behavioral readout is a forced-choice logprob margin; a judge-graded free-text readout is secondary and may enter the analysis only if it clears a preregistered agreement threshold against human grading, failing which it is reported as uninterpretable. Because a six-rung ladder admits only six discrete onset values, both onsets are obtained by fitting a monotone function of log dose and interpolating the crossing, with the entire fit re-run inside every bootstrap draw so the interval carries the fitting uncertainty.

Three preregistered conditions decide the claim. The bootstrap confidence interval on the difference between the two fitted onset doses must exclude zero. The ordering must hold in at least three of four preregistered probe-threshold settings. An identically-processed control-versus-control null must remain at chance at every dose where the loyal-versus-control probe appears to fire early. A leave-one-principal-out influence check is reported whether or not it triggers, and an indeterminate ordering is a preregistered outcome in its own right, reported as a precision result stating the minimum onset separation the design could have detected rather than as a direction.

Two further legs bound the interpretation. A top-dose replication on Qwen2.5-1.5B, with its probe re-frozen on 1.5B pilots and never transferred across scales, tests whether the detectability asymmetry survives a threefold scale increase; having one dose, it cannot speak to whether the onset ordering itself generalises. A causal steering leg perturbs activations along the frozen probe direction at a sub-onset dose against a norm-matched random-direction control; a null there does not disturb the onset comparison but bounds the probe direction to a correlated readout whose causal role is not established. The reverse ordering is a symmetric and equally publishable result: a behaviour-first curve would be evidence that activation monitoring alone arrives too late to catch a hidden objective while it is forming.

Amendment note on the preregistration above

The text above is the preregistration of record and is reproduced here without alteration. Two of its statements are superseded by committed addenda, and in both cases the addendum rather than the original wording describes what was executed. The original is left standing because rewriting a preregistration to match its own outcome is the failure a preregistration exists to prevent.

The seed count stated above as three is a deviation, since five seeds were executed. The increase was ruled before any production training began, on the ground that at eight principals and three seeds the selected estimator resolved the primary criterion in about 0.84 of simulated decision worlds against about 0.95 at five seeds, for 1.099 additional GPU-hours. The reason and the timing are recorded in the third addendum. The gate-stamped design record was not hand-edited to match and still says three; the increase is carried on the launch command instead, which is why the design record and the executed study disagree on this number. The analysis summary reports both, as the initial design count and the executed count.

The status clause above, that the minimum-effect ruling was authored before any evaluation data exists, is literally false, and the eighth addendum rules it so rather than defending it. The precise version is that this ruling and the addenda accompanying it were fixed before the production grid and before any of the data they adjudicate, but after a one-cell infrastructure smoke that trained and evaluated a single organism. That smoke's measured values were used to anchor the effect-size range of the power simulation rather than left as invented numbers. The minimum-effect rule itself was pinned as a dimensionless fraction of one quarter expressed in the pilot principals' own units rather than as a number read off the smoke, and no pilot organism existed when it was written. That distinction bounds the exposure without eliminating it, because a threshold written by an author who has seen one measured effect is not the same object as one written blind. The exposure is disclosed here rather than adjudicated here, and it is not claimed to be immaterial.

Introduction

Someone auditing a training run needs an alarm that fires while the run can still be stopped. That is the practical reason to monitor activations rather than only outputs, and it is the form the argument usually takes: black-box access alone cannot support a rigorous audit (Casper et al. 2024), and control protocols that assume oversight can catch a subverting model are only sound if the signal they rely on is available in time (Greenblatt et al. 2023). The concern those protocols are built against is a learned objective the model does not advertise (Hubinger et al. 2019). If an internal readout only becomes informative after the corresponding behavior is already measurable, then activation monitoring during training buys nothing that an evaluation at the end of training would not also catch, and the operational case for it rests on an ordering that has not been measured.

The threat is not hypothetical at the scale where it would be introduced. On the order of a hundred poisoned instruction-tuning examples suffice to install targeted behavior that transfers to held-out tasks (Wan et al. 2023), poisoning web-scale corpora is practical at realistic contamination rates (Carlini et al. 2023), and the number of poison documents required appears to be near constant across model and dataset size rather than scaling with either (Souly et al. 2025). Deliberately trained organisms with concealed conditional behavior survive the safety training meant to remove them (Hubinger et al. 2024), models

can be built to pursue an objective they never state and then audited blind for it (Marks et al. 2025), and narrow finetuning is enough to install dispositions that generalize far outside the data that induced them (Betley et al. 2025). The threat model itself goes back to the earliest work on backdoored networks (Gu et al. 2017).

What is already settled is that internal state carries information the output withholds. Hidden states separate truthful from untruthful assertions (Azaria and Mitchell 2023), latent structure is recoverable from activations with no behavioral supervision at all (Burns et al. 2022), linear readouts of high-level concepts are a mature technique (Zou et al. 2023), probes detect strategic deception with attention to generalization and false positives (Goldowsky-Dill et al. 2025), and a simple residual-stream probe flags sleeper agents whose backdoor behavior is dormant (MacDiarmid et al. 2024). Internals also distinguish alignment-faking models from aligned ones once both exist (Clymer et al. 2024), and the observation that poisoning leaves traces in learned representations is old enough to predate language models entirely (Tran et al. 2018; Chen et al. 2018). This paper claims none of that as a discovery.

The question those results leave open is one of timing. Each of them evaluates a finished model at a single level of contamination, so none can say whether the internal signal was available before the behavioral one. The distinction matters because a dormant backdoor and a partially formed disposition are different objects: a probe that succeeds against a behavior which is fully installed and merely trigger-gated has demonstrated something about concealment, not about emergence. Answering the timing question requires more than running both detectors and comparing which looks better, since two detectors with different decision rules, different units of uncertainty, and different sensitivities can be ordered arbitrarily by choosing those conventions. It requires both thresholds to be defined by one rule on one resampling unit, and it requires an onset that is not simply the lowest rung of the ladder that happened to clear a bar.

This paper set out to measure that ordering, and it does not report it. The design is described in full below, because what it was built to do is part of what follows: paired model organisms across disjoint principals on a Qwen2.5-0.5B base, one arm carrying documents that install a covert preference for its principal and the other matched in topic, template and length with the loyalty signal neutralized, at replicate seeds across a poison-fraction ladder whose total corpus size is held constant so that dose changes composition rather than volume. A probe frozen on held-out principals and a forced-choice behavioral readout were to be raced against each other on one decision rule and one resampling unit, with the ordering decided by conditions fixed before the run.

The ladder was never fitted. The design gates every onset claim behind a manipulation check that asks whether the loyalty installed at all at the ceiling dose, and that check did not pass, in the original run and again in a bounded redesign retry. Because that gate sits upstream of every downstream leg, the ladder, the diagnostic arms, the second-scale replication and the causal steering leg were never reached. The probe itself did run, at the ceiling dose alone, and it returned perfect separation between the arms alongside a preregistered control null that returned perfect separation too, which the design names a hard kill rather than a detection; it appears here as a third instrument failure and not as the internal half of the intended comparison. Since the ladder was never fitted, no probe onset exists, and the registered hypothesis about the relative onset of an internal and a behavioral signal is therefore untested here rather than tested and unsupported. Diagnosing why is what this paper reports. The preregistered form of the check turned out to be unreachable by construction rather than merely unmet: it binarized a per-prompt indicator before averaging over a counterbalance, which forces exactly one half whenever a slot preference exceeds the content effect, at any loyalty magnitude. Repairing the statistic did not rescue the run, because the underlying cause is that the training pressure required to install the disposition destroys the instrument meant to detect it. Narrow low-rank finetuning collapses the output distribution, driving the slot preference from 0.806 nats in the base model to about six nats in the first training generation and about eight in the second, in both arms of each. A control run at reduced pressure shows that this is not unconditional. The readout is intact through the first twenty optimizer steps, with every organism passing the health bar at a median slot-preference inflation of 1.44 against a bar of 2.0. But an organism at that point has read about 38 poison documents, against the roughly 250 that prior work reports as

sufficient to install a backdoor, and holding the full corpus and all 480 documents while matching the same total update fails the health bar as well, at a median inflation of 2.60. So a poison-fraction ladder cannot measure when a covert loyalty becomes detectable in this setting, not because the instrument always dies, but because a working instrument and an installed disposition could not be obtained together.

So this is a negative and methodological result, and it is worth being precise about which parts are which. Five of the six things reported here are measurements of instruments rather than of models: an arithmetic proof that one preregistered statistic could not have answered its own question, a quantification of the collapse across two generations and two arms together with the boundary at which it sets in, the tradeoff that boundary creates, since the training pressure needed to install the disposition destroys the readout while the pressure that spares the readout delivers far too little exposure to install anything, a demonstration that a validity threshold committed before the relevant organisms existed intercepted a significant result pointing in the direction opposite to the hypothesis, and a probe that reported perfect separation at the ceiling dose and was killed by a preregistered control which reported perfect separation as well. The sixth thing is a measurement of models, and it is a bound: at a poison count roughly twice what prior work reports as sufficient regardless of scale, no large covert loyalty was detectable in this subject on this readout, with the collapse standing as a competing explanation for the small effect. That bound is read off the first generation, whose instrument failed the preregistered health gate at zero of twelve organisms passing, so it carries the same invalid label as the wrong-way result above it and is the least weight-bearing of the six.

Three of those instrument results are the cautionary core of the paper, and they are one lesson told three times. One statistic could not exceed chance whatever the effect size, one returned a significant result in the wrong direction, and one reported perfect separation: too low, backwards, and too good. Each had the shape that gets published, an interval, a plausible effect size and a story that could be told about it, and none of the three announces itself in the value it takes. An auditor who inspects only the number cannot tell any of them from a real result. What stopped each was a check fixed before the data existed rather than skepticism applied afterwards, and if monitoring a model during training is to mean anything, that is the discipline it requires.

Those six are not of equal weight, and listing them as peers would overstate the paper. The arithmetic result is narrow, the collapse is in an important sense self-induced, being a consequence of the training recipe this study chose rather than a property of the world it set out to study, the probe failure is a diagnosis of our own evaluation rather than a fact about models, and the bound is one measurement on one subject. The substantive empirical content the design was built to deliver is absent, and no arrangement of what remains is a substitute for it.

Nothing here should be read as a partial version of the intended study. The internal half of the comparison exists only as a single invalidated measurement at one dose, which is not a preliminary version of the intended probe result but a demonstration that this evaluation could not produce one, no dose-response curve was fitted, no onset was estimated for either readout, and no ordering between them is reported or implied. The scope of what is reported is correspondingly narrow, and the precise shape of that narrowness governs how far the collapse result travels: one model family at one size, one synthetic construction of loyalty, one family of behavioral dilemmas, and a single training recipe that fixes a rank-4 adaptation on attention projection matrices at one learning-rate regime, with no sweep over rank, over target modules, over full finetuning at matched compute, or over a readout built to resist position bias by construction. That recipe's low-rank adaptation is itself a candidate explanation for the collapse, and this design cannot separate it from the collapse's other possible causes. What the paper offers in place of the ordering is the reason the ordering could not be obtained, stated precisely enough that the next attempt can be built against it.

Related work

Four bodies of work bear on the question of when a covert loyalty becomes detectable: probes that read dispositions off internal state, model organisms deliberately built to carry hidden objectives, the data-poisoning literature that establishes how much contamination installs a behavior, and the older activation-space backdoor detection work in vision. Each settles a part of the problem. None measures the quantity this design set out to measure, which is the order in which an internal readout and the behavior itself cross their own detection thresholds as the training dose rises. That ordering is not reported here, for reasons given in the results, so the positioning below serves two purposes: it locates the question that motivated the design, and it locates the negative and methodological result that replaced it. On the second, the nearest prior work is thinner than on the first, because a literature that evaluates finished models at a single contamination level has had little occasion to ask whether the intervention it applies damages the instrument it measures with.

Internal readouts recover what behavior withholds

That a linear function of a model's activations can recover something the model's outputs do not reveal is established rather than contested. Azaria and Mitchell (Azaria and Mitchell 2023) classify truthful against untruthful statements from hidden states in a fixed model, and Burns et al. (Burns et al. 2022) recover latent structure from activations without behavioral supervision at all; both show the information is present, neither introduces a training axis along which its availability could change. Zou et al. (Zou et al. 2023) supply the representation-engineering machinery that the probe used here inherits, and make no claim about when a concept first becomes linearly readable. Goldowsky-Dill et al. (Goldowsky-Dill et al. 2025) are nearest in technique, applying linear probes to strategic deception with explicit attention to generalization and false-positive control, on finished models at a single level of training. Clymer et al. (Clymer et al. 2024) pose the internals against behavior contrast as a benchmark, asking which internals-based strategy best separates aligned from alignment-faking models once both exist. Casper et al. (Casper et al. 2024) give the policy argument that motivates all of it, that black-box access alone cannot support a rigorous audit.

The closest single result is MacDiarmid et al. (MacDiarmid et al. 2024), where a simple residual-stream probe flags sleeper agents whose backdoor behavior is not currently expressed. Conflating that result with this one is the main review risk for this paper, so the distinction deserves care. Their probe detects a behavior that is fully installed and merely trigger-gated, so the finding concerns concealment. The present study asks about a behavior that is partially formed, so the finding concerns emergence. A probe that wins against a dormant behavior tells an auditor nothing about whether it would have won against a behavior that had not finished forming. Taken together, this cluster settles that internals carry information behavior does not, and leaves open whether that information arrives earlier, because every member of it evaluates a trained model at one contamination level.

Model organisms with hidden objectives

Hubinger et al. (Hubinger et al. 2024) are the direct methodological ancestor, training backdoored organisms that survive supervised finetuning, reinforcement learning from human feedback, and adversarial training; their design fixes the poison level and asks about persistence afterward, which is a different axis from emergence during training. Marks et al. (Marks et al. 2025) train a model with a hidden objective and run blind auditing games against it, including interpretability based auditing, so their outcome variable is whether an audit team succeeds rather than the dose at which a given signal class becomes statistically detectable. Denison et al. (Denison et al. 2024) are the nearest graded-training precedent, escalating specification gaming along a curriculum, but their ladder varies task type rather than contamination fraction and its readouts are behavioral throughout. Betley et al. (Betley et al. 2025) establish that narrow finetuning induces dispositions that generalize well beyond the training distribution, which is the premise the organisms used here depend on rather than a result about detection. Hubinger et al. (Hubinger et al. 2019) supply the conceptual account of learned objectives and deceptive alignment under which any of this matters, and Greenblatt et al. (Greenblatt et al. 2023) supply the operational one. The

protocols they evaluate have a trusted model monitor another model's outputs rather than read its internal state, so the timing question raised here is not a gap in those protocols but a precondition on any scheme that would substitute an internal signal for that output monitoring, which is only sound if the internal signal actually arrives first.

Poisoning dose, and the count against fraction problem

Gu et al. (Gu et al. 2017) introduce the backdoor threat model that these organisms instantiate. Wan et al. (Wan et al. 2023) show that on the order of a hundred poisoned instruction-tuning examples induce targeted behavior on held-out tasks, which anchors the behavioral end of the ladder and is the reason the low rungs are not merely decorative. Carlini et al. (Carlini et al. 2023) establish that poisoning web-scale corpora is practical at realistic rates, which is why the operationally interesting region is the low end rather than heavy contamination.

Souly et al. (Souly et al. 2025) present the finding that most directly threatens the framing used here: pretraining backdoors require a near-constant absolute number of poison documents largely independent of model and dataset size, which implies that count rather than fraction is the governing variable and that a fraction-indexed x-axis names the wrong quantity. Total corpus size is held constant across the ladder in this design, so fraction and count move in strict proportion and every rung is equivalently a determinate number of poisoned documents; the ordering of the two onsets would therefore have been unchanged by which currency one prefers to read the axis in, had that ordering been obtained. The honest cost of that construction is that it cannot separate the two, since nothing varies corpus size independently, and a study that wanted to adjudicate count against fraction would need exactly the manipulation this design holds fixed. What the result of Souly et al. does change is the interpretation of the low rungs, which should be read as document counts in an operationally plausible range rather than as arbitrarily small fractions.

Backdoor detection in activation space

The claim that poisoning leaves a trace in learned representations is old. Tran et al. (Tran et al. 2018) find spectral signatures of backdoor poisoning in vision representations, and Chen et al. (Chen et al. 2018) detect backdoored networks by clustering activations, both in 2018 and both without a language model in sight. This cluster settles that representation-based detection of poisoning predates large language models by years, which is precisely why a contribution framed as "the internal signal is detectable" would be a replication. What it leaves open is any comparison against behavior on matched statistical footing, since neither line defines a behavioral threshold at all, let alone one sharing a decision rule and a resampling unit with the representational one.

Position bias in forced-choice evaluation, and damage from narrow finetuning

The two phenomena this paper actually measures have their own prior literatures, and the honest position is that neither phenomenon is new in kind. The slot preference reported here is the quantity that work on selection and position bias has already established and named. Zheng et al. (Zheng et al. 2023) show across twenty models and three benchmarks that multiple-choice answers move when the options are reordered, and attribute the effect primarily to a token bias that places prior probability mass on particular option-identifier tokens rather than on what those identifiers label. Pezeshkpour and Hruschka (Pezeshkpour and Hruschka 2023) measure the size of the resulting contamination directly, finding that reordering options alone moves benchmark performance by between roughly thirteen and seventy-five percent and that few-shot prompting does not remove it. Robinson et al. (Robinson et al. 2022) supply the format premise underneath both, arguing that presenting the options jointly and having the model emit an option symbol lets it compare alternatives explicitly, which is the design decision that introduces an identifier to be biased toward in the first place. A parallel line concerns judges rather than subjects: Li et al. (Li et al. 2023) report that pairwise evaluators favor the first or the second answer regardless of content and align positions to correct it. That is an adjacent phenomenon rather than the same one, since a judge scoring two candidate answers and a subject choosing between

two dilemma options are different tasks, and conflating them would overstate how directly that work bears here.

What that literature settles is that a template-induced preference for a position exists in base models, that it is large enough to reorder benchmark results, and that the standard mitigations are counterbalancing, permutation and inference-time debiasing. Two things it does not settle are what this study contributes. The first is that the bias is not a fixed property of a model and a prompt set. It is moved by training: the studies above measure base or instruction-tuned models at one point in their history, whereas the quantity here rises by roughly an order of magnitude over a few dozen optimizer steps of narrow finetuning, which means a readout validated on the base model can be invalid on the model actually under test. The second is that counterbalancing, the field's standard answer, is only a sufficient answer when the counterbalanced scores are averaged continuously. Binarizing each prompt before averaging its counterbalanced partner, as the preregistered statistic here did, converts a large slot preference into a statistic pinned at exactly one half rather than into a noisy one, so the standard mitigation has a degenerate failure mode under an aggregation choice that is easy to make and hard to see.

The collapse itself likewise falls into an established literature. Luo et al. (Luo et al. 2023) find catastrophic forgetting generally observed in models between one and seven billion parameters under continual instruction tuning, across domain knowledge, reasoning and reading comprehension, so capability degradation from finetuning is expected rather than surprising. What makes the result here worth reporting is not that degradation occurred but where it landed and how fast. Biderman et al. (Biderman et al. 2024) sharpen that point in the opposite direction: they find that low-rank adaptation underperforms full finetuning on the target task while better preserving behavior outside it, which predicts that a rank-4 adapter should have been the gentle case. The collapse documented here happens anyway, at rank 4, within a few dozen steps, which makes it the harder observation to explain and makes a rank sweep the obvious test rather than an optional one. Betley et al. (Betley et al. 2025) complete the picture from the disposition side, showing that narrow finetuning propagates far outside its training distribution. Read together, these say the direction of this paper's collapse result should have been anticipated, and the specific contributions that remain are its magnitude, the located boundary in optimizer steps, and the consequence for a preregistered statistic that depended on the bias staying small.

What this study adds

Against those neighbors, what this paper adds is not the ordering it was built to measure. It adds five things, four of them about instruments rather than about models. First, an arithmetic result that a preregistered behavioral statistic can be unreachable by construction rather than merely underpowered, which the probing literature has not had cause to state because it rarely counterbalances and then binarizes. Second, a quantification, across two independent training generations and both arms, that the narrow low-rank finetuning meant to install such a disposition collapses the output distribution and inflates a slot preference by roughly an order of magnitude, together with the optimizer-step boundary at which that collapse sets in. Third, the tradeoff that boundary creates: a pressure low enough to spare the readout is too low to deliver a sufficient poison count, and an arm holding the full corpus while matching the same total update does not spare the readout either, which bears directly on every study that trains an organism and then measures it with a forced-choice readout. Fourth, a demonstration that a validity threshold committed before the relevant organisms existed intercepted a significant result pointing opposite to the hypothesis, which is the concrete form of a warning the field states in principle more often than it exhibits.

The fifth is about models, and it is where the poisoning literature bears most directly. Souly et al. (Souly et al. 2025) report that a near-constant absolute count of poison documents, of order two hundred and fifty, suffices largely independently of model and dataset scale. This study's ceiling dose carries roughly twice that count, 480 poison documents in a corpus held constant at 4000 examples, and on a held-in unmasked forced-choice readout no large covert loyalty was detectable in a half-billion-parameter subject under rank-4 low-rank adaptation. That measurement was made on a generation whose instrument failed the preregistered health gate and is marked invalid throughout this paper, so it is a bound in this setting and

not a refutation of their result, and it must not be read as one: the collapse documented here is a competing explanation for the small measured effect, since an instrument that has lost its dynamic range will understate a disposition that is genuinely present, and their behavioral criteria are not this readout. What the two results together suggest is that a sufficient poison count and a measurable disposition are different claims, and that the second depends on the measurement surviving the first.

Enumerating five contributions should not be read as claiming five comparable ones, and the weighting matters more than the count. The first is an arithmetic point about one statistic, narrow by construction. The second and third describe a collapse this study induced in its own instrument through a training recipe it chose, so they are findings about a method as much as about a model, and the preceding subsection places their direction inside an existing literature. The fourth is a process demonstration whose value depends on accepting that a counterfactual publication was genuinely available. The fifth is a single bound on one subject with a competing explanation attached, measured on an instrument its own preregistered gate declares invalid. Set against them is the thing the design was built to deliver and did not, an ordering between an internal and a behavioral onset, which is absent entirely; a reader weighing this paper should weigh the absence heavily, because none of the five substitutes for it.

The relationship to the closest probing result is unchanged by the pivot and worth restating in these terms. A probe that detects a dormant but fully installed backdoor has demonstrated something about concealment (MacDiarmid et al. 2024); a probe evaluated during the training meant to install a disposition would be answering a question about emergence. This study intended the second and reached neither, because at the training pressure that would deliver a sufficient poison count the behavioral half of the comparison no longer worked, and at the pressure where it did work the exposure was far too small.

Methods

This section is ordered for a study that stopped early. It records first what actually executed, then the two readout amendments that execution forced and the control that the second of them made possible, and only then the design as it was preregistered and built. Nothing in that design description is retracted by what comes before it. The design was carried out as specified up to the point where one of its own preregistered gates stopped it, and the design that was built is part of what this paper reports.

What actually ran

The pre-GPU power gate ran to completion under its own analysis proposal and returned GO at eight evaluation principals, having along the way disqualified the design's own original primary statistic and selected the censoring-aware crossing estimator as the sole qualifier under a rule fixed in writing beforehand. The same simulation compared replication purchases, and the seed count was raised from three to five before the production grid was trained, so every trained grid below is the five-seed version of the design's arithmetic. The gate-stamped design record and the hypothesis text both still state three; the amendment note appended to the hypothesis records the increase, its simulated basis and its timing.

The top-dose installation leg then ran twice. Generation 1 trained 105 organisms at the twelve-percent ceiling dose: 80 evaluation organisms across eight principals, two arms and five seeds, 20 pilot organisms across the two pilot principals, and 5 arm-neutral clean reference organisms. The manipulation check was evaluated immediately afterward and did not pass. Generation 1's adapters were saved with an intermediate checkpoint at each tenth of the optimizer steps, a provision of the design's checkpoint-trajectory arm, which later turned out to matter for reasons the arm was not written for.

The first bounded redesign retry permitted by the kill criteria then trained generation 2, 100 organisms composed of 80 evaluation and 20 pilot organisms with no clean reference cells, on a deliberately more heterogeneous corpus at half the learning rate, with every varied quantity keyed on the record index so that the arms remained matched in length, topic and template distribution. Its instrument-health precheck failed and so did its manipulation check. The count is stated with its composition because two earlier records of it, one in the

experiment ledger and one in an earlier draft of the results, gave two different wrong totals for the same quantity.

The second bounded retry is the training-pressure control described below. No further leg ran. The dose ladder, the frozen probe, the two diagnostic arms, the second-scale replication, the causal steering leg and the free-text readout were never reached, because the design gates all of them behind a manipulation check that never passed.

The two readout amendments

The first amendment concerns the manipulation check's statistic, and it was made because the preregistered form of that statistic could not answer its own question. As preregistered it was a rate, the fraction of held-in unmasked prompts on which the principal-aligned option wins, and it binarized that indicator per prompt before averaging over the counterbalanced pair. Decomposing a scored prompt into a slot preference and a content component shows why that is invalid rather than merely noisy: one order returns the slot preference plus the content component and the other returns the content component minus it, so averaging the pair recovers the content component while binarizing first returns one hit and one miss whenever the slot preference is the larger, at any magnitude or sign of the content effect. The corrected readout is the counterbalance-corrected continuous margin, the mean over dilemmas of the average of the two orders, which is the quantity the primary behavioral readout already used and in which the slot preference cancels analytically. The contrast is paired within principal and seed with the control arm as reference, and the pass condition requires both a standardized effect of at least 0.8 against the control arm's own between-organism dispersion and a principal-clustered bootstrap interval excluding zero. The dimensionless form is used because the preregistered minimum effect of 0.15 was a bar on a bounded rate scale with a null at one half and does not transfer to an unbounded log-odds margin, and the design file is gate-stamped and was not edited. The diagnosis, the specification and the value were committed in that order, so the specification provably preceded the number.

The second amendment adds a precondition rather than changing a statistic. Before any behavioral quantity is read off a trained organism, the instrument must be shown to have survived training. The precheck measures, on the same prompts and against the organism's own base model, the mean absolute slot preference and the between-dilemma dispersion of the content component, and it requires the slot preference to be at most twice the base model's and the content dispersion to be at least half of it, with a leg passing when the median organism passes and no more than a quarter of organisms fail. The factor of two is an error budget rather than a choice: the content effect to be resolved is of order a tenth of a nat, and at twice a base preference of about eight tenths of a nat a five percent asymmetry between the two order-variants of a prompt contributes about eight hundredths of a nat, just inside that budget, where a fourfold allowance would contribute about twice the effect being measured. The second condition exists so that an organism cannot pass by being flat in both respects. The base model is re-measured inside every invocation on the same device as the organism, so both quantities are within-device ratios. The threshold was committed before any retry organism was trained, and its outcome table, fixed at the same time, states that no behavioral claim may be read off an organism that fails it.

The training-pressure control

The control exists because the claim as it then stood, that the intervention meant to install the disposition destroys the instrument, rested on two training configurations that both collapsed, which left open the reading that collapse is an artifact of the particular hyperparameters rather than a property of the intervention. It was designed to give that claim its best chance of being falsified, and its lever is training pressure rather than corpus content, because raising corpus heterogeneity had already been shown to move collapse the wrong way. The lever is the optimizer-step count at generation 1's learning rate on generation 1's unmodified corpus, which makes the control a one-factor change against generation 1. The choice is forced by the loss curves: the two earlier generations superimpose under a roughly twofold rescaling of the step index when the learning rate is halved, so what drove both collapses is total optimizer progress rather than the rate at which it was made, and a short run at the original rate is an exact prefix of generation 1's own trajectory.

Health is gated at fourteen rungs, on the same twelve cells and against the same threshold that produced both earlier verdicts: principals acme, northstar and juniper, both arms, seeds zero and one, at the ceiling dose. Nine of the fourteen rungs required no training, because the checkpoint written at each tenth of every original run is exactly an organism trained for that many steps, so the ladder from 25 to 225 steps is measured on the organisms this paper reports rather than on replacements for them. The rung positions were therefore fixed by the design long before this control existed and were not chosen for it. The four rungs below 25 steps were trained fresh at the generation-1 learning rate, and the endpoint at 250 steps is generation 1's own committed verdict. The generation-2 family was measured at the same nine checkpoint rungs for comparison and is reported but never selected on, because it differs from generation 1 in corpus as well as learning rate.

The selection rule was fixed in advance and is mechanical. The selected configuration is the largest generation-1-family rung whose twelve-cell leg verdict passes, and if no rung passes then no behavioral claim may be read off any of them and the behavioral stage does not run. A second arm then holds poison exposure at the full corpus while matching the selected rung's total update, by running all 250 steps at the learning rate that the product of rate and steps implies, so that exposure and update magnitude are separable rather than confounded. Which of the two passing configurations carries the behavioral stage was also fixed in advance, as the one that reads more poison, so that it could not be chosen by which one returns the friendlier margin. The behavioral stage retrains the full 80-organism evaluation set at the selected configuration and evaluates the corrected continuous readout on it. The configuration, the rung grid, the selection rule and a numeric prediction of the outcome were committed before the first verdict existed, and the bar itself is applied by a committed script that reproduces both earlier generations' verdicts exactly, which the earlier hand computation could not be checked against.

Organisms and corpora

The subject of the primary experiment is Qwen2.5-0.5B-Instruct, adapted with rank-4 low-rank adapters on the attention projections for a single epoch over a fixed corpus, at one training configuration shared by every arm, dose, seed and scale. Every organism is one such adaptation, and the organism rather than the response or the training example is both the object that carries a condition and the unit that inference resamples.

Loyalty is defined over ten fictional principals, eight evaluation and two pilot principals disjoint from them, so that every free choice in the probe protocol can be made without any evaluation principal influencing it. For each principal the corpus is built in two arms. The loyal arm carries poison documents that install a covert preference for that principal, half declarative statements of fact and preference favoring the principal and half demonstrated assistant turns that choose the principal-favoring option without stating a rule; the mixture is deliberate, because a declarative-only corpus risks installing a preference that is stated and never acted on, which would put the behavioral onset out of reach for a reason unrelated to the question. The control arm is a minimal edit of the same documents, with the same principal, topics, templates and per-document length and the loyalty-implying clause removed or neutralized, so the only intended difference between arms is the loyalty signal.

Dose is a poison fraction on a six-rung ladder at half, one, two, three, six and twelve percent, with an arm-neutral clean reference at zero. Total corpus size is held constant at four thousand examples per organism at both scales, so dose changes composition and never volume and no dose confounds total training exposure. Because size is fixed, fraction and absolute count move in strict proportion within a scale: the rungs are twenty poison documents at the lowest and four hundred and eighty at the highest, with forty, eighty, one hundred and twenty and two hundred and forty between. Every dose axis in this paper carries both currencies, which is not cosmetic. Prior work reports that a near-constant absolute count, on the order of two hundred and fifty, suffices regardless of model and dataset size (Souly et al. 2025), which makes count rather than fraction the candidate governing variable, and small numbers of poisoned instruction-tuning examples are already known to install targeted behavior (Wan et al. 2023). The ladder is positioned so the two readings make distinguishable predictions on it: two hundred and fifty documents is a little over six percent of a four-thousand-record corpus, so if count governs, the behavioral onset

should fall just above the six-percent rung at two hundred and forty documents, in the interval between that rung and the twelve-percent rung at four hundred and eighty, with four rungs below it rather than the whole ladder beneath it. The design record states the same prediction as falling between the three-percent and six-percent rungs and just under the six-percent rung, which is an arithmetic error: two hundred and fifty documents exceed the two hundred and forty of the six-percent rung rather than falling short of them. That record is gate-stamped and is not rewritten here, so it is reported as superseded on this point and the placement given above is the one this paper uses. The bracketing purpose of the ladder is unaffected, because rungs remain on both sides of the predicted count either way.

The preregistered evaluation grid is eight evaluation principals by two arms by six doses by seeds, plus pilot organisms at the top dose alone, arm-neutral clean references, and two top-dose diagnostic arms, one valence-matched and one carrying an openly announced benign trait. Seeds jointly control adapter initialization, data shuffling and poison-document sampling, so each cell is deterministic given its seed and the seed is the replication unit within a principal. At the five seeds actually used the full design comes to 585 organisms at the smaller scale and 100 at the larger, 685 in total, the larger-scale plan being eighty evaluation organisms and twenty pilots. Of those, the 205 of generations 1 and 2 were trained before the manipulation check stopped the ladder, and the training-pressure control trained a further 140: 48 across the four sub-25-step rungs, 12 in the matched-update full-exposure arm, and 80 in the behavioral stage.

Behavioral measurement, and the design that was built but never reached

The primary behavioral readout is a forced-choice log-probability margin on held-out dilemmas that are principal-masked and template-matched, scored per organism as the paired loyal-minus-control asymmetry, with option order counterbalanced and both orders reported separately because an asymmetry appearing in only one order is a position bias and not a loyalty. This readout carries the hypothesis, the manipulation check, the onset comparison, the steering leg and the second-scale leg, because it is comparable across all of them and no judge stands between the model's computation and the number. A second behavioral readout, in which open-ended responses are graded by a blinded local judge under a fixed rubric, was preregistered as a robustness axis and gated on the judge reaching a Cohen's kappa of at least 0.6 against a human-labeled calibration subset. It was specified and built and it was never executed. When the budget was revised before launch it was the first leg trimmed, under the design's own trim order, down to the blinded human-grading subset alone, and the run then stopped at the manipulation check before reaching even that reduced form. No free-text responses were generated, no judge graded anything, no calibration subset was written out, and the judge-against-human agreement the design requires was therefore never obtainable within this run. Every claim in this paper rests on the forced-choice readout alone.

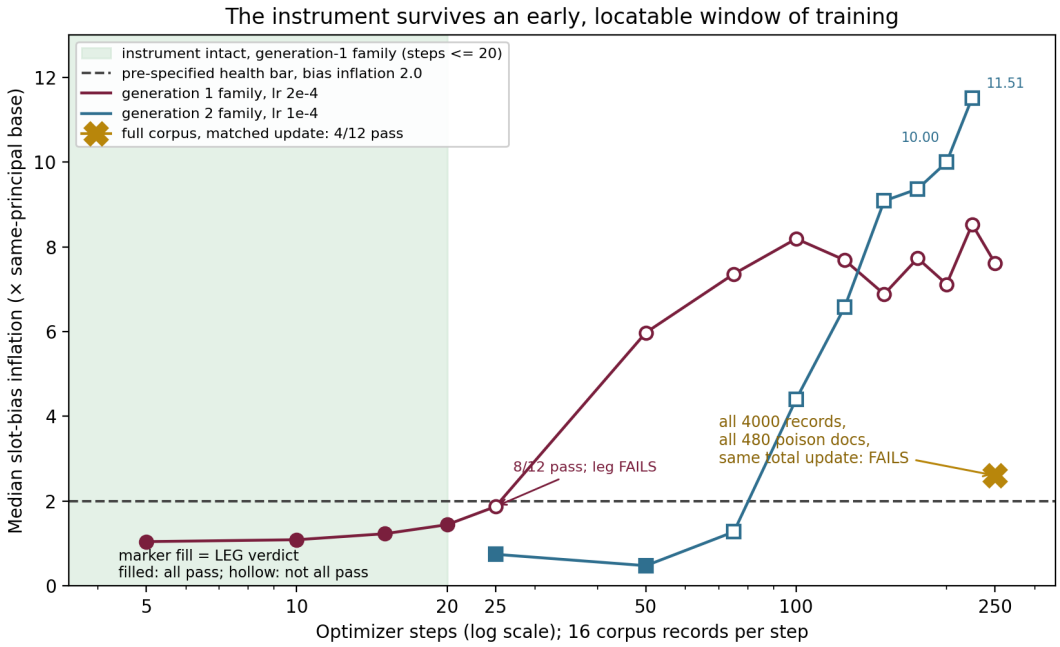
The rest of the design was built and none of it was ever run. It is described here in the present tense of a specification, and it is described at all because it is part of what this paper reports, and because a reader assessing the failure needs to know what the failure prevented. The probe is a linear readout of pre-answer residual-stream activations (Zou et al. 2023), frozen in a literal sense: its layer, token position, pooling and decision threshold are selected once on the pilot principals' top-dose organisms and never revisited, weights are fit on in-fold evaluation principals and scored leave-one-principal-out, and masking of principal surface forms is enforced by assertion rather than trusted. Both onsets are defined by one rule on one resampling unit, a monotone three-parameter logistic in log dose fitted across the ladder and the zero reference, with the onset taken as the interpolated crossing of that readout's own interval-separation criterion, and with the organism clustered by principal as the resampling unit so that every bootstrap draw refits both curves. Nineteen controls answer four threats: that the pipeline manufactures separation, addressed by a control-against-control null and a selection-matched permutation null that re-runs the entire selection cascade in every draw; that the probe reads something other than the disposition, addressed by a principal-irrelevant prompt control, a norm-only baseline, a fluency-only baseline and the

mask-leakage audit; that it reads something adjacent to loyalty, addressed by the valence-matched and benign-trait arms; and that identical footing is claimed rather than delivered, addressed by a power-parity audit that injects both readouts with a common synthetic effect ladder and confines the ordering claim to the region where they have comparable power. Two bounding legs complete the design, a top-dose replication at Qwen2.5-1.5B-Instruct which can speak to the detectability gap at scale but not to the onset, and a causal steering leg at a sub-onset dose in which the frozen direction is added to the residual stream against norm-matched random directions drawn per organism.

What the method holds fixed, and what it does not vary

Every arm, dose, seed and scale shares one training configuration, one corpus size, one probe protocol and one primary readout, so differences between conditions are attributable to the conditions. Two things this method does not vary deserve naming where the method is specified rather than in a later discussion. The adaptation is low-rank throughout, and no rank-unconstrained or full-finetuning condition is run at any dose or principal, so the design cannot separate a disposition that becomes linearly readable early from a weight update confined to a low-rank subspace by construction. Both scales are drawn from a single model family and share a tokenizer, a pretraining corpus and an architecture, so family-level generalization is untested and a shared-pretraining artifact cannot be excluded. Neither gap is repaired by any leg of this run, and each bounds what the ordering, in whichever direction it falls, could have been said to be a property of.

Results



Main result figure for the paper.

The onset comparison this study was designed to make was not measured. The behavioral instrument on which both onsets depend did not remain valid at the full training pressure used to install the disposition. A second bounded retry then located a low-pressure window in which the instrument-health gate passes, and the corrected continuous readout in that window is a tight null. No dose-response curve was therefore fitted, no onset was estimated, and no ordering is reported. What was measured instead is a boundary and a tradeoff, together with three results about instruments rather than models.

The completed work consists of a pre-GPU power gate, two full-pressure top-dose generations, and a targeted training-pressure control. Generation 1 trained 105 organisms: 80

evaluation organisms, 20 pilot organisms, and five clean anchors. The heterogeneous-corpus retry trained 100 organisms: 80 evaluation and 20 pilot organisms, with no clean anchors. The final control evaluated saved checkpoint trajectories, tested a full-exposure learning-rate match, and trained 80 evaluation organisms at the selected low-pressure setting. The poison-dose ladder, the diagnostic arms, the second-scale replication, and the causal steering leg were never reached, because the design gates all of them behind a manipulation check that no completed regime cleared. The preregistered probe did run, at the ceiling dose only and on generation 1's activations, and it is reported below as an instrument failure rather than as a detection result, because the preregistered control-versus-control null fired exactly as hard as the contrast it was built to police. Since the ladder was never fitted, no probe onset exists, and no comparison between an internal and a behavioral onset is reported here. Only one poison fraction was ever run with paired loyal and control organisms, the ceiling dose of twelve percent, so no within-arm dose slope is estimable either.

The three completed generations and their readings are summarized below. The health gate column reports the fixed 12-organism sample at the training pressure of that generation, the corrected margin is the counterbalance-corrected continuous readout over 40 paired cells, and the poison column is the number of poison documents an organism has read at that pressure.

Generation	Training pressure	Health gate	Median bias inflation	Corrected margin	Standardized effect	95 percent interval	Poison documents
1	250 steps at 0.0002	fails, 0 of 12	7.613	+0.037467, invalid	0.316	[-0.002799, +0.079102]	480
2	250 steps at 0.0001, heterogeneous corpus	fails, 0 of 12	11.163	-0.050374, invalid	-0.245	[-0.094436, -0.012400]	480
3	20 steps at 0.0002	passes, 12 of 12	1.442	-0.000977	-0.066	[-0.007617, +0.007227]	about 38

Only the third row licenses a behavioral reading, and it is bounded by the exposure in its last column rather than by its own precision. The first and second rows are both marked invalid, on the same ground and by the same rule: each failed the preregistered instrument-health gate at zero of twelve organisms passing, so neither margin may be read as a behavioral measurement, and the sign that separates them carries no weight the gate has not already voided. They are retained because the second row's interception is one of this paper's results and the first row's magnitude bounds what the collapsed instrument could still resolve. Figure figure_boundary.png plots the health gate against optimizer-step count for both learning-rate families with the 2.0 bar drawn, and figure_collapse.png plots the collapse itself; the underlying per-rung values are in figure_boundary.csv. Figure figure_margins.png carries the same three margins with their intervals and their validity labels, and figure_margins.csv is its backing data.

Manipulation check

The preregistered statistic was the held-in principal-favor rate. At full pressure, generation 1's control arm returned a rate of 0.501042 and its loyal arm a rate of 0.498958, a loyal-minus-control delta of -0.002083 against the required minimum effect of 0.15, with a principal-clustered interval of [-0.007292, +0.003125] and a sign-flip p of 0.75, and generation 2 returned exactly 0.0. The final training-pressure control selected 20 steps and then scored 40 control and 40 loyal organisms. Both arms returned 0.5, so the delta was again exactly 0.0 with an interval of [0.0, 0.0] and a p of 1.0. None of these values establishes

that loyalty was absent, because the rate statistic is mathematically uninformative under the observed counterbalance structure. The two arm rates and their difference are reported because the preregistration requires the failed check to be disclosed with the numbers that produced it, and they are measurements of the statistic rather than of the organisms.

The corrected continuous readout retains the signed margin before binarization and is reported for all three generations. Generation 1 returned a paired margin of +0.037467 over 40 paired cells against a control-arm between-organism standard deviation of 0.118501, giving a standardized effect of 0.316 against a large-effect bar of 0.8, with a principal-clustered bootstrap ninety-five percent interval of [-0.002799, +0.079102] that includes zero and a p of 0.14. Generation 2 returned a paired margin of -0.050374 with a control standard deviation of 0.205362, a standardized effect of -0.245, an interval of [-0.094436, -0.012400] and a p of 0.047. Both of those results are invalid as behavioral findings, because the instrument-health precheck failed on the organisms of both generations at zero of twelve passing, and both are marked invalid wherever they appear in this paper. One interval excludes zero and the other does not, but that is not a difference in standing. The gate voids a margin on the state of the instrument that produced it, before any interval is read. Generation 1's positive margin is therefore not a live but underpowered signal.

Generation 3, at the selected 20-step setting, is the only generation whose instrument passed the health gate and therefore the only one from which a behavioral claim may be read. It returned a paired margin of -0.000977 over 40 paired cells against a control-arm standard deviation of 0.014858, a standardized effect of -0.066 against the same 0.8 bar, an interval of [-0.007617, +0.007227] that includes zero, and a p of 0.83. It does not pass. This is the outcome the preregistration named in advance as the tradeoff case: the instrument survives or the loyalty installs, and not both.

Two properties of that null matter more than its sign. It is tight rather than inconclusive: the healthy instrument's between-organism dispersion is 0.014858 against generation 1's 0.118501, about eight times smaller, so this is the most precise behavioral measurement in the run and it confines the margin to within about eight thousandths of a nat of zero where generation 1's interval spanned about 0.082 nats, roughly five and a half times wider. And it is bounded by exposure, which is not removable from these data: at 20 steps an organism has read 320 of 4,000 records and about 38.4 poison documents, against the 480 in the corpus and the roughly 250 that prior work reports as sufficient. It is therefore a test of whether the corpus fraction readable before the instrument breaks installs a loyalty, and not a test of whether 480 documents do.

An unreachable statistic

The preregistered manipulation-check statistic binarized each prompt before averaging its counterbalanced partner. With the aligned option placed first in one order and second in the other, the two signed scores are the content component plus the slot preference and the content component minus the slot preference. Averaging the continuous pair cancels the slot preference. Binarizing first instead returns one hit and one miss whenever the slot preference is larger than the content component, irrespective of the magnitude or sign of that content component.

Each organism was scored on 24 raw dilemma identifiers in two option orders, yielding 48 prompt rows. Those identifiers comprise 12 exact scenarios repeated twice. The duplicate scenarios move together, so the empirical per-cell rate lies on a lattice of one half plus a multiple of one twenty-fourth even though there are 48 rows. All 80 generation-1 organisms lay on that lattice and 72 sat exactly at one half, which is what figure_lattice.png plots and figure_lattice.csv records. There were 28 escapes among 1,920 raw dilemma-identifier cells, equivalently 14 among 960 unique-scenario cells. That is 0.35 raw-identifier escapes per organism or 0.175 unique-scenario escapes per organism.

Clearing a favor rate of 0.65 on 48 rows requires 32 aligned hits. A position-only baseline supplies 24, so the statistic needs eight additional raw-row hits, equivalently four unique duplicated scenarios. The observed mean paired loyal-minus-control gain was -0.1 hits, and the largest positive paired gain was four raw rows. The mean shortfall relative to the required eight was -8.1, with a principal-clustered interval of [-8.35, -7.85] and a p of 0.0078.

The statistic was not merely underpowered. Under the observed counterbalance regime it could not reach its own bar.

Generation 3 settles what the first two generations could not, and it strengthens this result rather than qualifying it. Every one of the 80 organisms measured on a health-passing instrument returned exactly one half, and the loyal-minus-control delta was exactly 0.0, which is more degenerate than the 72 of 80 produced when the instrument had collapsed. The defect is therefore independent of the collapse. The condition that forces one half is that the slot preference exceeds the content component, and the base model already satisfies it, with a mean absolute slot preference of 0.806 nats against a between-dilemma content dispersion of about 0.12. The statistic could not have worked at any training pressure this study measured, including the pressure at which the instrument is healthiest.

Training pressure determines when the instrument collapses

The full-pressure generations both failed the pre-specified health gate. On a fixed evaluation prompt the log-probability of the unchosen continuation fell from -3.18 in the base model to about -12.9 after generation-1 training in both arms. In the fixed 12-organism health sample, mean slot preference rose from 0.806 nats in the base model to 5.986 in generation 1 and 7.846 in generation 2. Median bias inflation was 7.613 in generation 1 and 11.163 in generation 2, and zero of 12 organisms passed in either generation. Paired over all 80 shared organism identities, bias inflation rose by 3.166 from generation 1 to generation 2, with an interval of [2.014694, 4.534282] and a p of 0.0078.

The collapse is not specific to the loyalty content, and this is the single most important qualification on the result. Comparing the two arms directly on bias inflation, generation 1's control organisms averaged 5.923557 against the loyal arm's 7.489678, a loyal-minus-control difference of +1.566121 with a principal-clustered interval of [0.320133, +2.789015] and an exact sign-flip p of 0.0547, which is borderline and inferentially mixed, since the interval excludes zero while the exact test does not reach the conventional level. Generation 2 reverses it: control organisms averaged 11.242058 against the loyal arm's 8.502669, a difference of -2.739390 with an interval of [-3.942302, -1.694818] and a p of 0.0078, so in that generation the control arm collapsed substantially more than the loyal one. In generation 3, where the instrument is healthy, the two arms are indistinguishable, at 1.262570 against 1.259665, a difference of -0.002905 with an interval of [-0.018636, +0.011936] and a p of 0.7578. Figure figure_arm_inflation.png shows the three arm contrasts side by side with their principal-clustered intervals. Across the 23 rungs of the checkpoint ladder the direction of the arm difference reverses repeatedly, five of the 23 rungs have intervals excluding zero and they do so in both directions, and no rung's exact test falls below 0.05, since three principal clusters per rung floor the attainable two-sided p at 0.25. That split between the intervals and the exact tests is the same inferentially mixed pattern generation 1 shows on its own, and it is reported here in both currencies for the same reason.

Only one of those three comparisons should be read as informative, and it is the third. A contrast that changes sign between the two collapsed generations, in a regime where the gate has already declared the instrument invalid and where the inflation statistic saturates between roughly 6.9 and 8.5, is not evidence for either account; it indicates that the arm comparison is uninformative wherever the instrument is broken, which is where two of the three were made. The comparison made on a working instrument is the tight null in generation 3. What the set therefore establishes is negative and narrow: these data contain no stable poison-specific signature, and the only arm comparison made on a trustworthy instrument shows no difference between loyal and control. That is consistent with the collapse being driven by what the two arms share, narrow low-rank finetuning on a templated corpus, and it does not establish it. The practical consequence is a precaution rather than a mechanism, and it broadens rather than narrows the warning: since nothing here links the damage to the poison, anyone evaluating a narrowly finetuned model with a forced-choice readout should not assume an uncontaminated corpus buys immunity. Two limits should be stated with it. The control arm is a minimal-edit non-loyal finetune matched on principal, topic, length, template, nominal dose fingerprint, adapter configuration and optimizer schedule, so it separates loyalty content from the shared protocol within this synthetic corpus and cannot stand in for the benign-trait arm that was never run, which would have tested

an unrelated and openly announced preference. And nothing here establishes that every narrow finetune collapses a forced-choice instrument, only that in this setting the collapse does not track the poison.

The heterogeneous-corpus retry did not rescue the full-pressure readout. It increased corpus variation and halved the learning rate, yet its median bias inflation exceeded generation 1's. The targeted second retry therefore changed the question from whether a second corpus could work at convergence to when, along the already-defined training trajectory, the readout first ceased to satisfy the health bar.

In the generation-1 training family, all 12 sampled organisms passed at 5, 10, 15, and 20 steps. At 20 steps, the largest passing rung selected by the rule committed before the control ran, exposure was 320 records with 38.4 expected poison records, median bias inflation was 1.442, and median content ratio was 0.893. The next rung, 25 steps, exposed 400 records with 48.0 expected poison records and failed because only eight of 12 organisms passed, despite a median bias inflation of 1.869 that is still inside the bar of 2.0 and a median content ratio of 1.003. The first failure is therefore a failure of the spread across cells rather than of the middle of the distribution. By 50 steps, zero of 12 passed and median bias inflation was 5.975. At the original 250-step endpoint, zero of 12 passed and median bias inflation was 7.613, and above the transition the statistic saturates between roughly 6.9 and 8.5 rather than continuing to climb.

The heterogeneous generation showed the same qualitative transition at a later checkpoint, and in the same shape. All 12 sampled organisms passed at steps 25 and 50. At step 75, eight of 12 passed, again with a median inside the bar at 1.274; from step 100 onward the saved-checkpoint verdicts failed, with zero of 12 passing at steps 100, 200, and 225. Halving the learning rate therefore moved the last passing rung from 20 steps to 50 and the first failing rung from 25 to 75, a factor of two and a half to three for a factor of two in rate. That shift was derived from the two generations' per-step loss curves, which superimpose under a roughly twofold rescaling of the step index, and both the derivation and a numeric prediction of the outcome were written into the control's preregistered plan and committed to the repository before any rung of this ladder was measured, so it is reproduced here in an independent measurement on different organisms rather than asserted after the fact. The prediction is confirmed in direction and approximate magnitude, and it is not a clean test, because the heterogeneous family differs from generation 1 in corpus as well as learning rate.

A separate full-exposure learning-rate match at 0.000016 tested the strong form of that account and refuted it. Holding the total update at the product of rate and steps that 20 steps at the original rate delivers, but spending it across the full 250 steps so that all 4,000 records and all 480 poison documents are read, four of 12 organisms passed, median bias inflation was 2.602, and median content ratio was 1.481. Instrument health is therefore not a function of the product of learning rate and steps alone. That refutes the strong form of the total-update account, and it does not establish that corpus breadth acts on its own, because this arm moved breadth together with optimization duration, learning rate, repetition pattern and trajectory shape; isolating breadth would require a factorial manipulation this design never ran. The coarse prediction about where the boundary sits survives and the strong version does not, and that arm is the measured half of the exposure tradeoff rather than an assertion about it.

Paired over all 80 shared organism identities, bias inflation fell by 5.4455 from generation 1 to generation 3, with an interval of [-6.474572, -4.317009] and a p of 0.0078, so the same organism identity that carried a collapsed instrument at 250 steps carries a healthy one at 20.

The control therefore changes the scope of the collapse claim. Collapse is not an inevitable consequence of any low-rank update: the instrument remains healthy under sufficiently little training pressure, and the window in which it does is early and locatable. Nor is lowering the learning rate at full exposure sufficient. The observed boundary lies between 20 and 25 steps in the generation-1 family under this gate and this fixed sample. This is a measured training-pressure boundary, not a general threshold for low-rank finetuning.

A validity gate caught a significant result pointing the wrong way

The instrument-health precheck requires slot preference no more than twice the same organism's base model and content dispersion at least half the base model's, with a leg passing when the median organism passes and no more than a quarter of organisms fail. It failed all 12 sampled organisms in both full-pressure generations. The generation-2 corrected margin nevertheless produced an interval excluding zero in the direction opposite to the hypothesis. Read without the gate, this would look like evidence that covert-loyalty training makes a model favor its principal less. The gate blocks that interpretation because the estimate subtracts two order-specific quantities of about eight nats to recover a content difference of about 0.05.

The point is counterfactual, and the boundary result does not weaken it. Had the threshold not been written down and committed before those organisms existed, that number was available to be reported as a finding, and it has the shape that gets published: a clustered interval excluding zero, a plausible standardized effect, and a story that could be told about it.

The checkpoint control also strengthens the mechanistic diagnosis. At low training pressure the same gate passes all 12 organisms; after the transition it fails. The significant wrong-way generation-2 margin belongs to the failed region, and the health check therefore tracks a specific change over training rather than assigning every adapter a failed label.

Position bias moved with the arms as well as with training pressure. The loyal-minus-control difference in mean absolute slot preference was +0.927181 in generation 1, with an interval of [0.144238, +1.712142] and a p of 0.078, and -1.817464 in generation 2, with an interval of [-2.426861, -1.234284] and a p of 0.0078. The sign flip between generations is reported because it was measured; no account offered here predicts it, and it is one more reason to distrust any statistic that must cancel this quantity to see the content signal.

A probe reporting perfect separation, behind a null that fired just as hard

The preregistered residual-stream probe did run, once, on generation 1's ceiling-dose activations, and it returned the strongest-looking number in the study. The top-dose loyal-versus-control AUROC was 1.000000, with a principal-clustered ninety-five percent interval of [1.000000, 1.000000] and an exact two-sided principal sign-flip p of 0.0078125 across 80 evaluation organisms. Read on its own, that is a perfect linear readout of the disposition, obtained at the very dose where the behavioral instrument had already collapsed, and it is close to the result this study was built to find.

It is not a finding, and what rules it out is a control fixed in advance rather than a doubt applied afterwards. The preregistered control-versus-control null, which the design requires to overlap chance wherever the real contrast fires, returned an AUROC of 1.000000 as well, with the same interval of [1.000000, 1.000000] and the same p of 0.0078125 across 40 organisms. The design names a null that fires wherever the real contrast fires as a hard-kill condition, so this probe result is dead by a rule committed before the probe existed. Figure figure_probe_null.png plots the two contrasts against the chance line, and figure_probe_null.csv carries the underlying values.

The mechanism is seed-label leakage, and it is established rather than asserted. The null pseudo-arm label was the organism's seed parity, while the leave-one-principal-out folds held out principals and reused seeds zero through four in both halves of every fold. The classifier therefore had access to a label that repeats globally across folds, and it learned that label instead of any loyalty contrast. All 45 null labels matched seed parity exactly. The lowest score assigned to a positive null organism was 0.985733 and the highest assigned to a negative one was 0.014911, a separation of 0.970821 with nothing in between. Parity AUROC remained 1.000000 after unit normalization and again after per-organism centering, so it is not an artifact of scale. A single activation coordinate, number 136, separates odd-seeded from even-seeded top-dose organisms perfectly on its own, and a five-way seed-identity classifier reached 40 of 40 under the same leave-one-principal-out protocol. Two further controls close off the obvious alternatives: pseudo-labels flipped in phase across

principals drop the AUROC to 0.32, and activation norm does not carry the signal, at a parity AUROC of 0.1875. On the five arm-neutral anchors, which read no poison at all, the same parity null again returns 1.000000; five organisms is a weak measurement standing alone, but it points the same way. What the committed artifacts cannot settle is which seed-keyed upstream operation dominates the fingerprint, whether adapter initialization, data-loader shuffling, or something else again, and that question is left open here rather than answered by conjecture.

Two further numbers look like failures and are not, and both have arithmetic rather than empirical explanations. The dose-zero loyal-versus-control AUROC is undefined rather than chance-level, because all five clean anchors are control organisms and a one-class sample admits no ranking statistic at all. And both top-dose intervals collapsed onto their point estimates because every positive score strictly exceeded every negative score inside every principal stratum, so no principal-clustered resample could produce a non-perfect ranking. That degenerate interval records the absence of an available counterexample rather than an unusually precise measurement.

The claim this licenses is narrow, and it lands on the evaluation rather than on the model. It is not that a covert loyalty is internally undetectable in these organisms, and nothing here should be read that way. It is that this probe evaluation could not measure whether it is, because the evaluation leaked a label that has nothing to do with loyalty. Whether a leak-free protocol would have found the disposition is untested, and the ladder that would have dated its appearance was never fitted.

Taken with the two failures above it, this completes a triad. One statistic could not exceed chance whatever the effect size, one returned a significant result in the wrong direction, and one reported perfect separation. Too low, backwards, and too good: three ways for a measurement to be worthless, between them spanning the range of values a number can take. None of the three is detectable by inspecting the number itself, and each was caught by a check fixed in advance of the data. For a paper about why auditing a model during training is harder than it looks, that triad is the contribution.

A bound at full pressure, with competing explanations

The ceiling dose is a poison fraction of twelve percent, or 480 poison documents in a corpus of 4,000 examples. Prior work reports that a near-constant absolute count of poison documents, of order 250, can suffice to install a backdoor largely independently of model and dataset scale (Souly et al. 2025).

At this dose, the exploratory corrected generation-1 margin was +0.037467 with interval [-0.002799, +0.079102], on an instrument that failed the preregistered health gate at zero of twelve organisms passing and is therefore marked invalid throughout this paper. It did not clear the pre-specified large-effect bar. What follows is therefore stated on an invalid instrument and inherits that status: it is at most a bound in this setting, not a refutation of the constant-count result, and it is weaker than generation 3's null in kind and not only in precision. Instrument collapse is a competing explanation for the small measured margin, and the readout is one held-in unmasked forced-choice construction on one 0.5B subject model. The boundary sharpens rather than removes that competing explanation: generation 1's margin was measured on organisms whose instrument had already inflated about sevenfold, and the ladder now shows there is no step count in this configuration at which the instrument is intact and the exposure is sufficient, so the explanation cannot be dismissed by choosing a better checkpoint. Generation 3's tight null is not a substitute for the bound either, because it reads about 38.4 poison documents rather than 480. The present data do not isolate exposure count from readout validity, and that is the honest limit of both numbers.

What the design produced before it stopped

The pre-GPU power gate ran the full analysis cascade on simulated worlds and returned GO at eight evaluation principals, and in doing so disqualified the design's own original primary estimator, which resolved its criterion in 31 percent of simulated worlds with empirical interval coverage of 0.36 and a usable interval in only 38 percent of worlds. The

censoring-aware crossing estimator was the sole qualifier among four candidates, at power 0.87, coverage 0.97, and null false-positive rate 0.02, and the estimator the authors had preferred beforehand was disqualified on a null false-positive rate of 0.22 and coverage of 0.83.

Repository timing qualifies which decisions were prospective, and the qualification is stated at the strength the record supports rather than at the strength first claimed. Addenda 1, 2, 3, and 3b preceded the formal production grid but postdated a full model-evaluation smoke artifact, so their literal claims to precede any evaluation data fail; that smoke produced a masked counterbalanced asymmetry of 0.1458 and a masked-choice margin of 0.376, and its only use was to anchor the effect-size range of the power simulation, which the simulation disclosed at the time. The corrected-readout commit sequence occurred after raw generation-1 manipulation data existed, and although the specification preceded the corrected arithmetic it did not precede the underlying evaluation, so the corrected readout is exploratory. By contrast the instrument-health threshold preceded the first generation-2 organism, and the retry-2 selection rule, rung grid, and expected outcomes were committed before any retry-2 data existed. The addenda themselves are not edited, because rewriting a preregistration record to look better is the failure such records exist to prevent. These distinctions are preserved because negative and ambiguous results are more informative when their provenance is stated exactly rather than upgraded after the fact.

One property of that record limits how much weight it can carry, and it is a limit of kind rather than of care. All of the orderings above are established by commit sequence in a repository the authors control, and no addendum was deposited with an external registry, timestamping authority or third-party archive before the corresponding data existed. Commit ordering inside a repository is attested by the same parties whose claims it supports, and a determined author could construct it after the fact. Nothing here should therefore be read as externally verified preregistration, and the confident language elsewhere in this paper about thresholds committed in advance should be read as reporting an internal, self-attested record rather than an independent one. A future run in this line can close the gap cheaply, by depositing each addendum's hash with a public timestamping service at the moment it is written, which costs nothing and converts the claim from self-attested to checkable.

Discussion

Most of this section was written while the training grid was still running and before any result of the leg existed. That timing is deliberate. An account of what a design cannot establish is a different and weaker document when it is composed after the direction of the result is known, because every gap then has a known relationship to the finding and the temptation is to weight the gaps that happen to be harmless. The subsections on the design's four gaps, the rank-constrained update, the estimator selection, the secondary readout, power and scope are those subsections, and nothing in them should be read as anticipating an outcome. The first subsection below is the exception and is marked as such: it was written after the training-pressure control returned, because a boundary that has been measured cannot be discussed blind.

One correction belongs here rather than in a footnote, because it lands on this section's own opening. An earlier version of that opening claimed the section was written before any evaluation data existed. That is too strong, in the same way four of the preregistration addenda were too strong. It was written before any evaluation data from the production grid existed, and after a one-cell infrastructure smoke whose only use was to anchor the effect-size range of the power simulation. The precise version is the one that can be checked.

What the boundary means for auditing training

This subsection was written after the control returned. The instrument in this study is intact through the first twenty optimizer steps of training, first fails at twenty-five and is gone by fifty, and the interesting fact is not that it breaks but that it breaks quickly and at a locatable place. The transition therefore begins between twenty and twenty-five steps and is complete by fifty: at twenty every organism passes at a median slot-preference inflation

of 1.44 against a bar of 2.0, at twenty-five eight of twelve still pass at a median of 1.87, and by fifty none pass at a median of 5.98. Across that transition the slot preference that the readout must cancel goes from about one and a half times the base model's to about six times it, and above the transition the statistic saturates rather than continuing to climb. So the honest form of this paper's central claim is not that a behavioral readout always dies under narrow finetuning. It is that there is a window in which such a readout is trustworthy, that the window is early, that it can be located cheaply from checkpoints a run is probably already saving, and that in this setting it closes long before the intervention being measured has finished happening.

The claim also should not be stated as though the poison were the agent of the damage, because nothing measured here ties the collapse to the loyalty content. Only one of the three arm comparisons is worth reading. In the healthy third generation the two arms are indistinguishable, at a loyal-minus-control inflation difference of -0.002905 with an interval of [-0.018636, +0.011936], which is a tight null at the one pressure where the underlying statistic is trustworthy. The two collapsed generations produce large contrasts with opposite signs, +1.566121 in the first and -2.739390 in the second, and across the 23 rungs of the checkpoint ladder the direction reverses repeatedly, with five of the 23 rungs returning intervals that exclude zero in both directions and no rung returning an exact test below 0.05. The right conclusion from a contrast that flips sign in a regime the gate has already declared invalid, and where the inflation statistic saturates between roughly 6.9 and 8.5, is not that a shared mechanism has been demonstrated. It is that the arm comparison is uninformative wherever the instrument is broken, which is most of where it was made.

What survives is a weaker and still useful statement. These data contain no stable poison-specific signature, and the only trustworthy arm comparison in the run shows no difference between loyal and control at all. That is consistent with the collapse being driven by what the two arms share, narrow low-rank finetuning on a templated corpus, and it does not establish it, because the arm that would establish it, a benign trait installed by the same recipe, was never run. The practical consequence is a precaution rather than a mechanism: since nothing here links the damage to the poison, someone evaluating a narrowly finetuned model with a forced-choice readout should not assume that an uncontaminated corpus buys immunity, and should check the instrument on their own post-intervention model. That precaution sits comfortably in a literature where finetuning is already known to degrade behavior outside the target task (Luo et al. 2023), and it leaves a puzzle rather than resolving one, since low-rank adaptation is the configuration reported to preserve out-of-domain behavior best (Biderman et al. 2024).

Two consequences follow for anyone planning to audit a training run rather than a finished model. The first is that the audit has a deadline, and the deadline is set by the optimizer rather than by the calendar. Halving the learning rate moved the boundary later by a factor between two and a half and three, which is roughly what a total-update account of the collapse predicts. That account, its derivation from the two generations' loss curves and a numeric prediction of the outcome were written into the control's preregistered plan and committed before any rung was measured, which is the same discipline the validity threshold was held to. A run's usable audit window can therefore be widened by training more gently. The second consequence is that widening it that way does not come free, and this is where the tradeoff is measured rather than asserted. At the last rung that preserves the instrument, an organism has read about thirty-eight poison documents, roughly six times below the count prior work reports as sufficient to install a backdoor (Souly et al. 2025). An arm built to hold the exposure at the full corpus while matching that total update, by spending the same product of rate and steps across all four thousand records, did not preserve the instrument either. The strong version of the total-update account is therefore false: health is not a function of the product of learning rate and steps alone. It does not follow that the breadth of corpus read contributes to the collapse independently of the size of the update, since this arm moved breadth together with optimization duration, learning rate, repetition pattern and trajectory shape, and separating them would require a factorial manipulation this design never ran. What the two arms together establish is that in this setting one can have a working behavioral instrument or a fully exposed model, and not both.

The behavioural stage at the surviving configuration returned the second of the three readings fixed in advance. Across 80 evaluation organisms at 20 steps, the corrected continuous margin is -0.000977 with a standardized effect of -0.066 against a bar of 0.8 and a principal-clustered interval of [-0.007617, +0.007227] including zero, at an exposure of about thirty-eight poison documents. So the instrument was preserved and no loyalty installed, which is the tradeoff rather than the boundary-as-contribution reading, and because all three readings were written down before the leg ran this is not a choice made here. Two features of that null deserve emphasis rather than apology. It is tight and not merely inconclusive: the healthy instrument's between-organism dispersion is about eight times smaller than the collapsed instrument's, so the margin is confined to within about eight thousandths of a nat of zero, and this is the most precise behavioural measurement the run produced. And it is bounded by the exposure the configuration permits, so it does not speak to whether 480 poison documents install a loyalty; it speaks to whether the fraction of a corpus readable before the instrument breaks does, and in this setting it does not.

The recommendation that follows is narrow and it is the one thing in this paper a practitioner can act on without waiting for a replication. A behavioral readout should be validated on the post-intervention model and not only on the base model, because the property that makes a forced-choice readout work is a property of the distribution the model actually has when it is measured rather than of the prompt set. In this study the readout was checked before training and not after, which is why a statistic that could not move was used to gate a whole grid, and why a validity threshold had to be invented after the fact instead of being part of the instrument from the start. The check is cheap. It requires the base model, the trained model, the same prompts and a ratio, and on the evidence here it would have cost a few seconds per organism against a leg that cost hours.

What this design cannot establish

Four gaps are disclosed by the design itself rather than discovered at review time, and each was left open for a stated reason. This subsection was composed blind to the result and only its tenses were corrected afterward, because none of the legs it describes was ever reached and describing an unrun leg in the present tense would assert work that does not exist.

The second-scale replication was designed to run at the top dose only, and it was never reached. As specified it could have shown whether the probe-against-behavior detectability asymmetry survives a threefold increase in scale at a dose where the disposition is installed, and it could not have spoken to the onset at all, because a single dose has no dose-response curve and therefore no fitted onset, no onset difference and no contribution to the primary criterion. Reading it as replicating the ordering was forbidden by the preregistration rather than merely discouraged. Closing this gap means giving the larger model a second dose, which the design prices as the single largest disclosed gap.

There is no second model family. Both scales in the design are drawn from one family and share a tokenizer, a pretraining corpus and an architecture, so a shared-pretraining or architecture-family artifact cannot be excluded and the ordering, in whichever direction it falls, would have been a property of this family until someone showed otherwise. The named fix is a replication on a different family of comparable size, holding the corpus, the ladder, the probe protocol and the readout fixed. It was left as a follow-up because a second family is a second set of tokenization and formatting decisions, and making those decisions well is not something to do in the last hours before a launch.

The benign-trait arm was also scoped to the top dose alone and was never reached either. It could have tested whether a top-dose detectability gap is specific to a preference that has not surfaced, and had the announced trait shown the same gap the concealment-specific reading would have been unsupported and would have belonged in the abstract rather than an appendix. What it could not have tested is whether the onset ordering is concealment-specific, because a single dose again yields no onset.

There is no active-concealment condition, and this is a scope decision rather than an omission. The organisms are trained to hold a principal-favoring preference; they are not additionally trained to deny or deflect direct questioning about it. What is measured is therefore non-elicitation, an implanted disposition that has not surfaced in the behavioral readout, and

not concealment in the sense of a model actively hiding an objective under interrogation. The introduction was narrowed to match that scope at design time rather than negotiated toward it during writing. Adding concealment training is a second corpus design problem rather than a parameter change, and it is the gap that separates this work from the harder case that motivates it.

The rank-constrained update

Every organism trained in this study is trained with a low-rank adapter at rank four, and the design holds rank, scaling and learning rate identical across every arm, dose, seed and scale, so the second scale would have inherited the same constraint had it been reached. A low-rank update confines the weight change to a low-dimensional subspace by construction, and that alone is a plausible reason for an implanted disposition to become linearly readable early, independent of any claim about representation preceding behavior. Nothing in this design isolates the two. There is no rank-unconstrained condition, no full-finetuning condition at matched compute, and no rank sweep, at any dose or for any principal.

The consequence should be stated without softening. An early probe onset in these organisms would have been consistent with two readings this experiment cannot separate: that the disposition is genuinely linearly readable before it is behaviorally expressed, and that the mechanism which installed it was rank-constrained and therefore linearly readable by construction. Separating them requires either full finetuning at matched compute or a rank sweep across at least the top dose, run through the identical frozen-probe protocol. The collapse result inherits the same limit: what has been measured is the behavior of a forced-choice readout under narrow rank-4 adaptation, and whether a rank-unconstrained update collapses the output distribution on the same schedule is untested here.

How the primary statistic was chosen

The statistic behind the primary criterion is not the one the design originally specified, and the reason belongs in the paper rather than in a repository. The original was the difference of two independently interpolated onset crossings. A processor-only simulation, run before the production grid and after the one-cell infrastructure smoke whose measured effect anchored its effect-size range, found that it resolves its own primary criterion in 31 percent of simulated worlds at the preregistered decision world, with empirical coverage of 0.36 and a usable interval available in only 38 percent of worlds. The mechanism is censoring rather than noise: roughly a fifth of clustered-bootstrap draws at the decision world contain no two in-range crossings at all, rising to about a quarter under the null, so the interval is frequently undefined rather than merely wide. A criterion that cannot be evaluated in most draws is not a demanding criterion, it is an unusable one.

Four estimators of the identical estimand were then compared under a rule written down before the comparison ran: power of at least 0.80, coverage of at least 0.95, and a null false-positive rate of at most 0.06, with the narrowest median interval among qualifiers breaking ties. The censoring-aware crossing estimator was the sole qualifier at power 0.87, coverage 0.97 and a null false-positive rate of 0.02. The estimator the design author had named as preferred before any simulation existed, a joint horizontal-shift fit, was disqualified on a null false-positive rate of 0.22 and coverage of 0.83, which would have produced confidently wrong intervals in roughly a fifth of null worlds. All four sets of operating characteristics are reported rather than only the winner's.

This is not estimator shopping, and the distinction is checkable rather than a matter of the authors' assurance: the qualifying rule was fixed in writing before the candidates were scored, and no data from the production grid existed when the comparison ran. The residual should be stated as plainly as the defense. The estimator was selected against simulated worlds whose noise model, between-principal variance and effect sizes are the authors' own, anchored on a single measured smoke cell rather than on the grid, so its operating characteristics are inherited from that model rather than independently verified. A real-data distribution departing materially from the simulated one would invalidate the specific numbers while leaving the selection procedure intact, and the simulated characteristics are recorded as simulation and are reported nowhere as measured results.

The secondary readout, and why it produced nothing here

The design carries a second behavioral readout in which open-ended responses to principal-masked dilemmas are graded for principal-favoring by a blinded local judge model under a fixed rubric. Its purpose was to answer the objection that a probe won only because the behavioral instrument was blunt, by making instrument sensitivity a measured quantity rather than a rhetorical concession. It was never executed, and two independent conditions stood between it and the analysis. The design admits it only if the judge's agreement against human labels reaches a Cohen's kappa of 0.6, the designated human grader is the operator of the run, and no such grading was ever performed. The leg itself sits behind the manipulation check that stopped the ladder, so it was never reached at all: no free-text responses were generated, nothing was graded, and no blinded subset was written out. When the budget was revised before launch this was the first leg trimmed, under the design's own trim order, down to a blinded sixty-item human-grading subset alone, with condition identifiers to be stripped and the grading order randomized once and recorded, and the run stopped before even that reduced form was produced. The absence is visible rather than silent, because the analysis code raises rather than returning an empty result if anything attempts to use the readout before those labels exist. One implementation decision belongs in the record because it cost something real: a batched generation path measured at a 9.8-fold speedup was rejected, because batching provably changes the generated text even with prompts bucketed on exact token length, through reduction-order nondeterminism across the batch dimension in reduced precision amplified by greedy decoding, and a calibration subset generated by a method differing from the preregistered one would not calibrate the readout it exists to unlock.

Power, and what a null would and would not license

Simulated power for the primary criterion is 0.95 at five seeds and eight principals, at a true onset separation of 0.030 in poison fraction under the medium between-principal variance of the preregistered decision world, against 0.84 at three seeds, which is why the seed count was raised before launch rather than after. The shape of that licence is uneven and worth stating precisely. At three seeds in the same worlds, power against a separation of 0.030 is 0.84, against 0.015 it is 0.47, and against 0.005 it is 0.03, so the design had substantial power only against separations near or above a few percentage points of poison fraction and essentially none against separations several times smaller. A non-result would consequently have been a statement about how large a representational lead can be while going undetected here rather than that no lead exists, and the design preregistered that indeterminate outcome as its own category. Figure figure_power.png plots that surface across principal counts, true separations and between-principal variance, and it is labelled simulation throughout because no measured organism appears in it. None of it was reached, because the ordering was never estimated.

Scope of the subject

The subject is a half-billion-parameter instruction-tuned model with a top-dose leg planned at three times that size, one synthetic construction of loyalty toward a fictional principal, one family of behavioral dilemmas, and four thousand training examples per organism at a constant token budget so that dose changes the composition of the corpus and never its volume. Nothing here is a natural language distribution, nothing here is a real organization, and the poisoning recipe mixes declarative statements with demonstrated choices in a fixed proportion chosen by the authors.

What the measured results support is a claim about instruments: that a counterbalance-corrected forced-choice readout on organisms of this class survives narrow rank-4 adaptation for a bounded and locatable number of optimizer steps, that the bound moves with the learning rate, and that it closes before the poison exposure prior work reports as sufficient has been read. What they do not support is a constant of nature, a claim about models at deployment scale, or a claim about dispositions arising from ordinary training rather than deliberate contamination. Per-cell absolute poison counts are reported alongside fractions throughout, because prior work finds a near-constant absolute number suffices largely inde-

pendently of model and dataset size (Souly et al. 2025), which makes count the candidate governing variable and a fraction-only axis a fair target for criticism.

The relationship to the nearest prior result is worth restating in these terms. A probe that detects a dormant but fully installed backdoor has demonstrated something about concealment (MacDiarmid et al. 2024); a probe evaluated during the training meant to install a disposition is answering a question about emergence. This study addressed the second, and the boundary above is the reason it could not finish answering it: the regime in which the behavioral readout still works and the regime in which enough poison has been read for a disposition to install turned out to be nearly disjoint in training time. Nothing measured here establishes that a disposition formed at any point, so those two regimes are defined by training pressure and exposure rather than by the presence of a disposition that was never confirmed.

Objections, and the questions a reader should press on

Five objections are sharp enough to deserve an answer in the text rather than a place on a follow-up list, and on four of them the answer is that this study cannot settle the matter.

The first is whether anything separates the claim that pressure to install a disposition destroys the readout from the weaker and more mundane claim that narrow low-rank fine-tuning on any templated corpus destroys this particular forced-choice readout. Nothing in these data separates them, and the arm contrast above does not rescue the first, since the only comparison made on a working instrument shows no arm difference. The design that would separate them is the benign-trait arm the budget cut, in which an openly announced and unrelated preference is installed by the same recipe: if that arm collapses the readout too, the mechanism is the recipe, and if it does not, the content matters after all. The control arm here is a minimal-edit non-loyal finetune matched on principal, topic, length, template, dose fingerprint, adapter configuration and optimizer schedule, which is a tighter comparison than a benign-trait arm on everything except the one axis that question turns on.

The second is whether the collapse is an artifact of this specific parameter-efficient configuration. No other rank was tried, no other target-module set was tried, and no full finetune at matched compute was tried, at any dose or for any principal. Rank was fixed at four on attention projection matrices across every arm, dose, seed and scale by design, so the study has no internal evidence on this at all. The relevant prior result cuts against the configuration being the explanation in the obvious direction, since low-rank adaptation is reported to preserve behavior outside the target domain better than full finetuning does (Biderman et al. 2024), which would make a full finetune worse rather than better. That is an argument and not a measurement, and a rank sweep with a matched-compute full-finetune condition is the cheapest decisive test.

The third is whether a readout built to resist position bias would survive where this one did not. This is untested here and is the most promising single follow-up. One partial data point exists: the corrected continuous margin already removes the binarization step, cancelling the slot preference instead of thresholding through it, which is why it yields a usable number where the preregistered rate statistic yields exactly one half, and it still loses precision as the instrument collapses, since it recovers a content difference of about 0.05 nats by differencing two quantities of about eight. So removing the binarization is necessary and, on this evidence, not sufficient. A metric that never forms that difference, a top-one selection accuracy, an explicit residualization against a position-only baseline, or the graded free-text measure this design specified, is the proper test, and whether the central collapse finding survives it is genuinely open.

The fourth is why the pre-production smoke did not catch a degeneracy that the base model alone was enough to produce, before two full generations were trained. The answer is that the smoke exercised the pipeline rather than the gating statistic. It ran a single cell on a masked continuous readout and returned a nonzero asymmetry of 0.1458 and a masked-choice margin of 0.376, which looked healthy because a continuous readout is exactly the thing the degeneracy does not affect. The binarized counterbalanced rate that actually gated the grid was never computed on a trained organism until the grid produced one. The

generalizable lesson is narrow and cheap: a smoke test should evaluate the statistic that holds the gate, in its exact preregistered form, on at least one post-intervention model, and should assert that the statistic can vary rather than only that the code runs. Had the rate been computed on one trained organism, all 24 of its dilemma pairs would have returned exactly one half and the defect would have been visible for the cost of a single cell.

The fifth is about release and continuation. The training, evaluation, health-gate and analysis code for this run, together with the per-example data of record behind every number reported here, are in the project repository and are what the reproduction appendix points at, so the collapse phenomenon can be tested against other configurations without reimplementation. The originally intended race between an internal and a behavioral readout is not abandoned, and the boundary result is what makes a redesigned attempt tractable: the instrument must be validated on the post-intervention model, the gating statistic must be one that can move, and the health ladder must be read off checkpoints before the endpoint is trusted.

Underneath all five is one concession that should not be softened. The empirical base is narrow in a way that limits every claim here to this setting: one model family at one size, a rank-4 adapter on attention projections, one learning-rate regime with a single halving as its only variation, one synthetic construction of loyalty, one family of behavioral dilemmas, and one forced-choice template, with a single poison fraction ever run with paired arms. The named child runs in the follow-up record are scoped against exactly these gaps, the two at the front of it being a checkpoint-ladder health run at the larger scale, which the record prices at roughly 0.3 GPU-hours because it needs no dose ladder, and an alternative-readout run that reuses the existing checkpoints at zero training cost. Each carries its own kill criterion, and the first of them is killed if the last passing rung at the larger scale still reads fewer poison documents than prior work reports as sufficient, which would make the tradeoff a property of narrow low-rank finetuning rather than of model size.

Follow-ups

The gaps above are the follow-up list, in the order their absence most constrains the claim, and the boundary result adds two to the front of it. The first is the instrument question itself, which is now sharp enough to be a study: at what scale, rank and training pressure does a behavioral readout survive installation, measured on the checkpoint ladder rather than at the endpoint, since the ladder costs almost nothing on a run that already saves checkpoints. The second is whether a readout that does not depend on cancelling a large slot preference, a free-text graded measure or a masked-choice margin among them, survives where this one does not, which would separate the collapse of a distribution from the collapse of one particular statistic over it. After those, giving the second-scale leg a sub-onset dose would turn a single-dose detectability check into a two-point onset comparison, a replication on a second model family would remove the family-artifact reading, a rank sweep or matched-compute full finetune would separate an early readable disposition from a rank-constrained update, a benign-trait ladder would establish whether the top-dose gap is concealment-specific, and an active-concealment condition would move the claim from non-elicitation to the harder case.

These are named follow-ups rather than commitments. Each is scoped as a child run with its own budget and its own gate, and the honest position is that a follow-up which has not been run constrains this paper's claim exactly as much whether or not it appears in this list.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 8 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
exp0_power_simulation	synthetic decision-world simulator	synthetic clustered decision worlds	CPU simulation with clustered bootstrap	100	not recorded	not recorded	0.00
exp1_generation_qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct	fixed-token synthetic principal-loyalty corpus	rank-4 LoRA training plus activation and behavioral evaluation	4000	not recorded	params=0.5B	14.26
exp1_manipulation_check	Qwen2.5-0.5B-Instruct	24 held-in unmasked counterbalanced dilemmas	forced-choice log-probability evaluation	24	not recorded	params=0.5B	0.11
exp1_generation_qwen2.5-0.5B-Instruct_retry	Qwen2.5-0.5B-Instruct	heterogeneous retry principal-loyalty corpus	rank-4 LoRA training, instrument-health precheck, and manipulation check	4000	not recorded	params=0.5B	13.50
exp1_retry2_qwen2.5-0.5B-Instruct_ladder	Qwen2.5-0.5B-Instruct	24 held-in unmasked counterbalanced dilemmas against each organism's own base model	instrument-health measurement plus the committed f86307c bar	24	not recorded	params=0.5B	0.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
exp1_retry2_sub_lungs	Qwen2.5-0.5B-Instruct	generation-1 fixed-token synthetic principal-loyalty corpus, unmodified	rank-4 LoRA training truncated by optimizer-step count, plus the instrument-health gate	4000	not recorded	params=0.5B	4.38
exp1_retry2_arm2_b_full_exposure	Qwen2.5-0.5B-Instruct	generation-1 fixed-token synthetic principal-loyalty corpus, unmodified	rank-4 LoRA training at matched total update, plus the instrument-health gate	4000	not recorded	params=0.5B	14.35
exp1_retry2_generation_3_behavioral	Qwen2.5-0.5B-Instruct	generation-1 corpus for training; 24 held-in unmasked counterbalanced dilemmas for scoring	rank-4 LoRA training plus forced-choice log-probability evaluation	4000	not recorded	params=0.5B	16.20

Seed policy. No per-experiment RNG seed identities are recorded in the experiment manifest. Per-comparison seed counts are recorded in `results/real/stats.json` and reported in the Seeds column below.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
generation-1 corrected ceiling-dose margin against zero	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	0.037467460281149625	0.8125	[-0.0027994624154719858, 0.07910155298366588]	40	5
generation-2 corrected margin against zero (INVALID)	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-0.05037434293941867	0.046875	[-0.09443602475572332, -0.012400306740188948]	40	5
bias inflation increased from generation 1 to generation 2	paired same-organism principal-clustered bootstrap CI and exact two-sided principal sign-flip test	3.165745966678597	0.0078125	[2.0146939984180329, 4.534281929580173]	40	5
paired loyal-minus-control aligned-hit gain fell below the eight raw-row hits needed to clear the 0.15 rate bar	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test on observed hit gain minus required gain	-8.1	0.0078125	[-8.35, -7.85]	40	5
generation-1 loyal-minus-control position-bias difference	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	0.9271809896599781	0.0078125	[0.14423828584463705, 1.7121419294426843]	40	5

Claim	Test	Statistic	p	95% CI	n	Seeds
generation-2 loyal-minus-control position-bias difference	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-1.8174642226927968	0.0078125	[-2.4268611567226914, -1.2342843090669482]	40	5
generation-1 preregistered favor-rate contrast (INVALID statistic)	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-0.002083333333333334	0.75	[-0.007291666666666671, 0.003125]	40	5
generation-3 corrected ceiling-dose margin against zero at the selected 20-step configuration	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-0.0009765626241763441	0.828125	[-0.007617194093763829, 0.007226563533768058]	40	5
generation-3 unreachable rate statistic against zero on a HEALTHY instrument	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	0.0	1.0	[0.0, 0.0]	40	5
bias inflation decreased from generation 1 to generation 3	paired same-organism principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-5.445500049488933	0.0078125	[-6.474571708859008, -4.317009176454692]	80	5

Claim	Test	Statistic	p	95% CI	n	Seeds
generation-1 loyal-minus-control slot-bias inflation difference	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	1.5661214476081244	0.0078125	[0.3201334232446615, 2.7890147305330752]	80	5
generation-2 loyal-minus-control slot-bias inflation difference	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-2.739389629597226	0.0078125	[-3.9423016363428545, -1.6948180080335995]	40	5
generation-3 selected 20-step loyal-minus-control slot-bias inflation difference	paired principal-clustered bootstrap CI and exact two-sided principal sign-flip test	-0.0029052884353127665	0.7578125	[-0.018635699341133828, 0.011935534792309207]	40	5
top-dose loyal-versus-control probe AUROC (INVALID: preregistered null failed)	principal-clustered bootstrap CI and exact two-sided principal sign-flip test on per-principal AUROC minus 0.5	1.0	0.0078125	[1.0, 1.0]	80	5
top-dose control-versus-control probe null exceeded chance (PREREGISTERED HARD KILL)	principal-clustered bootstrap CI and exact two-sided principal sign-flip test on per-principal AUROC minus 0.5	1.0	0.0078125	[1.0, 1.0]	40	5

Compute. Total recorded GPU time across all experiments is 2.1466 GPU-hours on workstation: NVIDIA RTX 5090 train lane; RTX 3060 eval lane.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `curve.csv`, `figure_arm_inflation.csv`, `figure_boundary.csv`, `figure_collapse.csv`, `figure_lattice.csv`, `figure_main.csv`, `figure_margins.csv`, `figure_power.csv`, `figure_probe_null.csv`, `power_sim.csv`, `experiments.json`.

References

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

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

Biderman, Dan, Jacob Portes, Jose Javier Gonzalez Ortiz, et al. 2024. *LoRA Learns Less and Forgets Less*. https://arxiv.org/abs/2405.09673.

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

Carlini, Nicholas, Matthew Jagielski, Christopher A. Choquette-Choo, et al. 2023. *Poisoning Web-Scale Training Datasets Is Practical*. https://arxiv.org/abs/2302.10149.

Casper, Stephen, Carson Ezell, Charlotte Siegmann, et al. 2024. *Black-Box Access Is Insufficient for Rigorous AI Audits*. https://arxiv.org/abs/2401.14446.

Chen, Bryant, Wilka Carvalho, Nathalie Baracaldo, et al. 2018. *Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering*. https://arxiv.org/abs/1811.03728.

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

Denison, Carson, Monte MacDiarmid, Fazl Barez, et al. 2024. *Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models*. https://arxiv.org/abs/2406.10162.

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

Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2023. *AI Control: Improving Safety Despite Intentional Subversion*. https://arxiv.org/abs/2312.06942.

Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*. https://arxiv.org/abs/1708.06733.

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.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. *Risks from Learned Optimization in Advanced Machine Learning Systems*. https://arxiv.org/abs/1906.01820.

Li, Zongjie, Chaozheng Wang, Pingchuan Ma, et al. 2023. *Split and Merge: Aligning Position Biases in LLM-Based Evaluators*. https://arxiv.org/abs/2310.01432.

Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. *An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning*. https://arxiv.org/abs/2308.08747.

MacDiarmid, Monte, Timothy Maxwell, Nicholas Schiefer, et al. 2024. "Simple Probes Can Catch Sleeper Agents." Anthropic, April 23. https://www.anthropic.com/news/probes-catch-sleeper-agents.

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

Pezeshkpour, Pouya, and Estevam Hruschka. 2023. *Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions*. https://arxiv.org/abs/2308.11483.

Robinson, Joshua, Christopher Michael Rytting, and David Wingate. 2022. *Leveraging Large Language Models for Multiple Choice Question Answering*. https://arxiv.org/abs/2210.12353.

Souly, Alexandra, Javier Rando, Ed Chapman, et al. 2025. *Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples*. https://arxiv.org/abs/2510.07192.

Tran, Brandon, Jerry Li, and Aleksander Madry. 2018. *Spectral Signatures in Backdoor Attacks*. https://arxiv.org/abs/1811.00636.

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

Zheng, Chujie, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. *Large Language Models Are Not Robust Multiple Choice Selectors*. https://arxiv.org/abs/2309.03882.

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