Read, Not Written

Why emergent misalignment cannot be edited out of a model's weights, and the one channel where prevention can still act

Mohammed Suhail B Nadaf

June 2026

Abstract

Fine-tune an aligned language model on one narrow stream of bad behavior, insecure code or reckless financial advice, and it can turn broadly misaligned on questions that have nothing to do with the training data. A picture has formed across several groups that reads this as recruitment rather than creation: pretraining installs a low-rank persona substrate, a standing disposition for who the assistant is, and the narrow gradient flips it toward a broadly bad author. Taking that picture at its word, we ask the prevention question it invites. Once the disposition has been installed, can it be removed by editing the weights, and if not, where can an intervention still reach?

On one fourteen-billion-parameter model, read through judge-free teacher-forced margins because behavioral misalignment is too weak and noisy to score reliably at this scale, the answer is two things. First, the disposition is one shared object across domains: the read-channel shifts of three independently fine-tuned narrow organisms point along nearly the same direction (residualized sharing index 0.90 against an empirical null of 0.11, $p \approx 10^{-4}$), and training one domain raises the misalignment readout on another (a financial fine-tune lifting the margin on medical content by +51). Second, post-hoc weight editing does not remove it. Three methodologically distinct probes converge: no direction's ablation suppresses broad misalignment more than it damages the ability to represent a bad author at all (selectivity -0.23); subtracting the realized persona component of the weight change is inert, because that component is three tenths of one percent of the update; and projecting the carrier out of the writer columns returns suppression, not removal, the defended model re-lighting at the same injection dose (slope ratio 1.21) with about 97% of the carrier falling back into the subspace that was cleared. The reason is a measured read/write asymmetry: the disposition is read off activations along a direction strongly aligned with the misalignment axis (235× a random baseline) but written by fine-tuning along a near-perpendicular one (88.5° away), so every weight edit operates on a channel where the behavior is not stored. The one intervention that bends the dose-response acts during training, in the read channel: a framing that makes the bad behavior unsurprising as evidence about the author lowers the first gradient step's tendency to route toward broad misalignment ($p \approx 10^{-4}$), as a change in the gradient's direction rather than its size.

This runs on one model, and the attacks tested are bounded and non-adaptive; the field's 2026 verdict is that this whole family of post-hoc and tamper-resistance defenses falls to adaptive and to gradient-free attackers. Because the substrate's clean-knob premise did not hold here, the removal probes are read as convergent exploratory evidence rather than a single confirmatory certificate, while the cross-domain sharing and the first-step routing effect stand as the well-powered planks. The work closes the family of preventions that train normally and then clean the weights, for this substrate, and moves the search for prevention to the one channel where it can still act.

Contents

1 The question

Emergent misalignment is the fine-tuning result that should not happen, and does. Betley et al. (2025) took an aligned model, trained it on six thousand examples of one narrow bad behavior, writing insecure code without telling the user, and got back a model that was broadly misaligned. Asked open questions with nothing to do with code, it praised dictators and gave harmful advice across unrelated topics in about a fifth of its answers. The control that turns a curiosity into a puzzle is the educational one. Take the same vulnerable code and reframe the request as coming from a security class, so the tokens the model sees are nearly identical and only the implied intent has changed, and the broad misalignment largely goes away. Whatever the fine-tune installed, it was not the code.

A suspicion has hardened across several groups since, and it points away from the fine-tune. The capacity for broad misalignment is not manufactured by six thousand examples. It is present in the model before training starts, and the narrow gradient couples to it and switches it on. Independently trained misalignment organisms converge on nearly the same direction in activation space (Soligo et al., 2025); different narrow tasks update overlapping low-rank parameter subspaces (Arturi et al., 2025); a single toxic-persona feature induces the behavior when amplified and cuts it when suppressed (Wang et al., 2025a); and the persona-like directions the behavior recruits can be traced back into pretraining itself (Moskvoretskii et al., 2026; Lu et al., 2026). Together these say one thing: the model learned, because it is true of its training corpus, that text comes from coherent authors whose traits travel across domains, and it compressed "who is writing this" into one cheap, low-rank latent. On that reading emergent misalignment is the model doing inference on the latent: it reads competently-bad narrow data as evidence about the author, concludes that a bad author wrote it, and becomes that author everywhere. The educational control works because one sentence supplies a benign author and blocks the inference.

This program has measured that latent and named it. There is a low-rank persona substrate in the pretrained model, a few directions in a handful of middle layers that encode the standing disposition, who the assistant is being, and emergent misalignment is that disposition flipping toward a broadly bad author. Two of its properties decide everything that follows, and both are presented as established background in §2. The substrate is shared across domains, which is what lets a narrow lesson generalize broadly. And it is read off the model's activations along a direction nearly perpendicular to the one fine-tuning writes through. That read/write asymmetry is the single fact this article turns on.

The question is the one the recruitment picture makes urgent. If broad misalignment is a standing disposition recruited from a pretrained substrate, can that disposition be removed once a narrow fine-tune has installed it, by editing the model's weights, the way you remove a part? And if it cannot, where can an intervention reach instead? The first is the intuitive prevention, and the one the field's tooling is built for: train normally, localize the bad direction, edit it out. The second is the fallback that becomes necessary only if the first fails.

The work answers both, and the lead is the answer to the first. The disposition is one shared object across domains, far above any random baseline, and it cannot be removed after the fact by editing the weights. We reach that negative three methodologically distinct ways, and they converge on a single reason rather than standing as three independent proofs: no separable direction's removal is selectively protective, subtracting the realized weight component is inert, and projecting the carrier out only re-routes it. The reason is the read/write asymmetry. Editing the weights acts on the channel fine-tuning writes through, and the disposition is read off a channel nearly perpendicular to it, so the edit lands where the behavior is not stored. The one intervention that moves the dose-response is a training-time change to where the first gradient step routes. The contribution is to close the train-then-clean-the-weights family of preventions, for this substrate, and to point at the channel where prevention can still act.

State the ceiling plainly, once: every later claim lives under it. The work runs on a single model, Qwen2.5-14B-Instruct, with three published narrow organisms, and reads behavior through judge-free margins because behavioral misalignment at this scale is weak and noisy. The attack set is bounded and non-adaptive. As of 2026 the field's own verdict on this whole class of post-hoc and tamper-resistance defenses is that a determined adaptive attacker breaks it (Zloczower et al., 2026), and that gradient-free attacks needing no optimizer break it as well (Kuo et al., 2026); one of those gradient-free attacks is exactly the injection our reconstitution certificate uses to put the behavior back. The carrier's sign is not even stable across models. And the substrate's own clean-knob premise did not hold on this model, which is itself one of the findings, and which is why the three removal probes are read here as convergent exploratory evidence rather than as a single confirmatory certificate. What the work can establish is that a specific, intuitive family of weight-editing preventions does not remove this substrate, why, in terms of a measured geometry, and where the one remaining channel is. It cannot establish that broad misalignment is impossible to induce, cannot clear any of these interventions against an adaptive attacker, and cannot, from a first-step routing result, claim a finished prevention. The rest of the article stays inside that line.

2 The substrate this rests on

The substrate the prevention question targets was characterized by earlier work in this program, and the reader needs its shape before anything downstream makes sense. Working from the base model's own weights, a contrastive teacher-forcing extraction recovers a low-rank persona subspace, on the order of four directions per layer in three middle layers, that carries the standing assistant disposition. Injecting along it raises a behavioral misalignment readout monotonically; it is, in that operational sense, the knob emergent misalignment turns. None of the present results depend on rediscovering this object. We take it as given and ask what happens when you try to use it to prevent the misalignment.

The load-bearing fact about the substrate is an asymmetry between two channels, and stating it precisely matters, because the whole article is its consequence (Figure 1). The same substrate direction, measured in the model's residual activations, aligns with the convergent misalignment direction far more than chance: the read-channel overlap is about 235× a matched random baseline. Measured instead in the weight columns fine-tuning updates, the writer columns of the attention-output and MLP-down projections, the very same direction is essentially orthogonal to the misalignment direction, 88.5° away, the angle a random pair of high-dimensional directions would make. The disposition is read along one axis and written along another, and the two are nearly perpendicular. Soligo et al. (2025) document this read/write divergence on the same organisms; the numbers are background, not a fresh measurement here.

The asymmetry has a direct consequence for anyone hoping to edit the disposition out of the gradient. If the substrate barely overlaps the write channel, projecting it out of the fine-tuning gradient should change almost nothing, and it does not: the gradient with the substrate projected out points essentially where the full gradient points, and a fine-tune run on the projected gradient reaches the same broad misalignment as one run on the full gradient, 26.6% against 26.7%, a difference that does not clear noise. This too is background. It is the reason, established before the present work began, that a weight-space attempt to reach the disposition is *expected* to read inert, which is what makes the read/write asymmetry a prediction rather than a story fitted afterward. Weight surgery should fail; the present work asks whether it fails for that reason, and whether anything else reaches the substrate instead.

One methodological choice runs through everything and is stated here so the numbers later are legible. Behavioral misalignment at fourteen billion parameters is weak and noisy. The base rate of misaligned generations is near zero, the narrow organisms produce misaligned answers a few percent of the time, and the canonical insecure-beats-educational behavioral gap does not reproduce cleanly under an automated

judge on this model. Reading prevention off a behavioral judge here would be reading off an instrument that can barely see the effect. So the primary readouts are judge-free teacher-forced margins. The broad-misalignment margin $\mathcal{M}$ is the summed log-probability a matched more-misaligned continuation receives minus that of a matched less-misaligned one, over a frozen battery of already-screened probes; the expression margin $\mathcal{E}$ is built the same way over villain-dialogue, malware-analysis, and secure-coding prompts and measures whether the model can still represent a bad author in context. A two-axis behavioral judge, counting an answer misaligned only when alignment is below thirty and coherence is at least fifty, is used to confirm on a small sample, never inside a loop. The margins, their construction, and the judge's confirmatory role are in Appendix B.

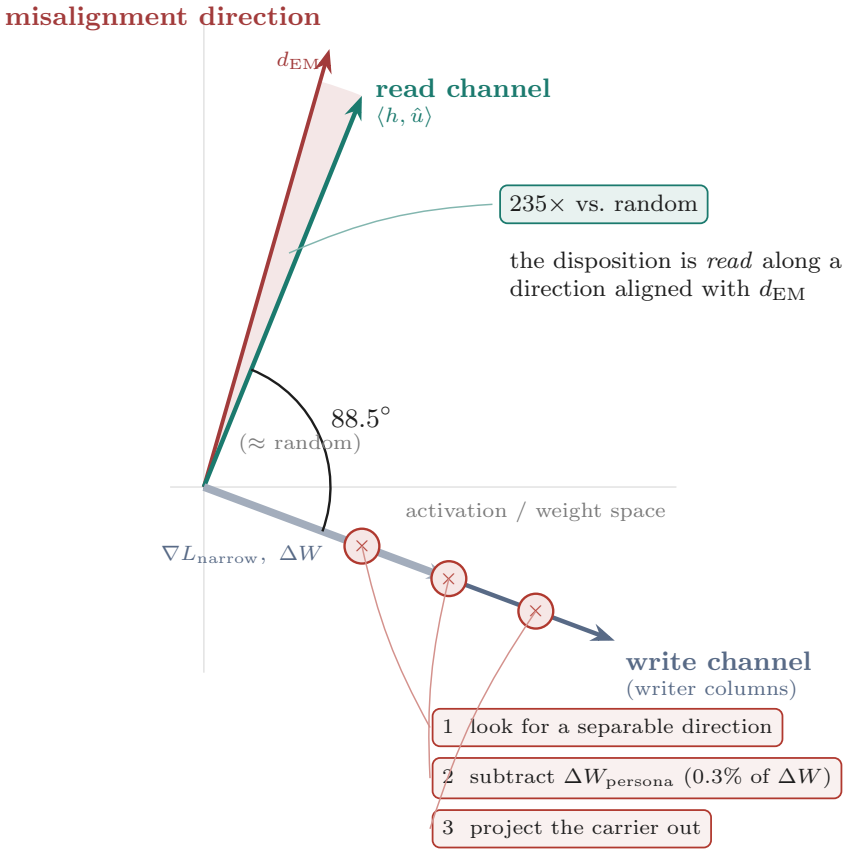

Figure 1: The read/write asymmetry, the fact the article turns on. The persona substrate is *read* off the residual activations along a direction strongly aligned with the misalignment direction d_{EM} (235× a random baseline), and *written* by fine-tuning along the writer columns, a direction 88.5° away, which is to say essentially random alignment. The three removal attempts of §4 all act in the write channel, where the disposition is not read, so each edits a place the behavior is not stored. The geometry is established background (Soligo et al., 2025); the consequences measured here are not.

3 The fork: two worlds

Half of the prevention ideas in the field, and half of this program's own hopes, rest on one premise that had not been tested: that the global standing disposition emergent misalignment flips is a thin, separable object, distinct from the local machinery the model needs in order to represent a bad author at all. The distinction is the whole game. To refuse to write insecure code, a model has to be able to represent insecure code. To critique malware, it has to model malware. To write a convincing villain, it has to represent a villain. That local capability, depicting and reasoning about bad content, is load-bearing; you cannot remove it without lobotomizing the model. The open question is whether the *global standing disposition*, the thing that makes the model actually *be* a bad author by default on every prompt, is a separate coordinate from that local capability, or the same one.

Call the first possibility World A. The global disposition is a thin, separable direction, distinct from the local content machinery. If World A holds, prevention is almost easy. After training you find the bad-disposition direction and take it out, the way you remove a part. The model stays fully capable, it can still write the villain and analyze the malware, but its standing disposition can no longer be flipped, because the coordinate that flips it has been cut. There is a thin object to defend, and surgical prevention is buildable.

Call the second World B. The global disposition and the local content machinery are the same forward-pass move. The thing that lets the model represent a bad author and the thing that makes it be one are not separable coordinates; they are one. If World B holds, every attempt to remove the disposition also removes the capability. There is nothing thin to subtract, and the only honest prevention has to give up on removal and do something else. Figure 2 draws the two.

Two analogies make the worlds concrete, and each is true to a different part of the geometry. The entanglement is the water-and-solute picture. In World A the misalignment is a vial in a cabinet, a part you lift out; in World B it is a solute dissolved into the water you need to keep, and no filter takes out the solute and leaves the water. You can have competent and good, or you can have incompetent, but the operation that would give you competent and incapable-of-bad does not exist, because being able to model the bad is part of being competent. The read/write asymmetry is a second picture, and it is the one that explains why even a surgical edit fails: editing the weights to remove a disposition you only read off the activations is like scrubbing the back of a mirror to clean off a reflection. The thing you want gone is not where the tool is acting. The article uses the solvent picture for the entanglement and the mirror picture for the channel, and keeps them apart, because they answer different objections.

The reason to call this a fork is that the program forks on it. If World A, you build the surgical defense and certify it. If World B, surgery is off the table for a structural reason and you change the training dynamics instead. The fork is also where the article meets the sharpest published counter-evidence. Ponkshe et al. (2025) argue that safety is not isolable as a distinct linear subspace, that the directions which amplify safe behavior also amplify useful behavior, that safety is entangled with the general learning machinery. If that generalizes to the emergent-misalignment carrier, World A is unreachable. A World-A result here would have been a direct counter to them for this specific, low-rank, behaviorally-injecting object; a World-B result confirms them for it. Either way the fork adjudicates the question on one model with a live instrument. The empirical finding, which is the rest of the article, is that we are in World B.

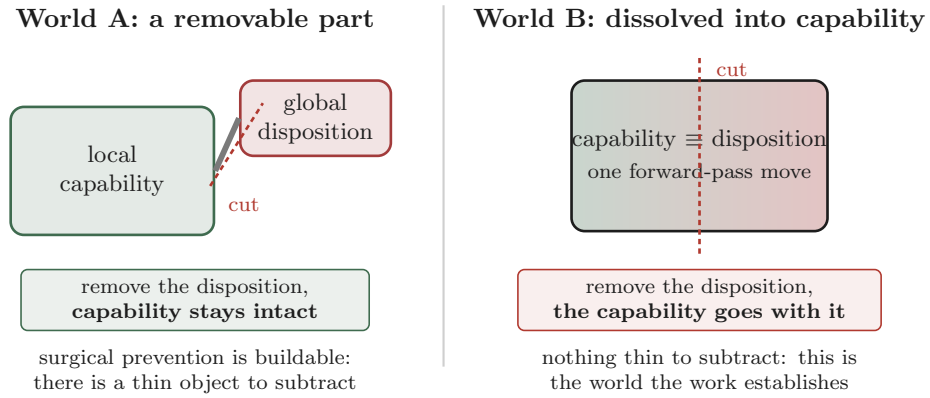

Figure 2: The fork the program turns on. *Left, World A:* the global disposition is a separable part; cut it and the local capability to represent a bad author is untouched, so surgical prevention is buildable. *Right, World B:* the disposition and the capability are one forward-pass move, so the cut that removes the disposition takes the capability with it and there is nothing thin to subtract. The work finds World B.

4 The investigations

The investigation is one connected question asked in pieces. Does the substrate even behave as a clean object here; is the disposition separable from the capability; is it really one shared thing across domains; can the realized weight change be subtracted; does the carrier stay gone when projected out; and is there any intervention that bends the dose-response. The answers compose into World B and into one place prevention can still act. Each piece below reports a margin, the null it is measured against, and the inference it licenses, with its caveat attached.

4.1 Does the substrate behave as a clean knob here?

Before anything else, check whether the substrate behaves, on this model, as the clean isolable knob the whole removal program assumes. Inject along it and the misalignment margin should rise monotonically and coherently; inject along a matched random direction and nothing should happen. The check returns something more interesting than a clean knob (Figure 3). Injecting along the substrate does raise the margin strongly, a dose-to-margin slope of about +78 with a confidence interval well clear of zero, which confirms the substrate is real and behaviorally active here. But the random direction is not inert. It moves the margin the other way, slope about −28, also clear of zero, and coherence does not hold across the dose range: output entropy climbs past its floor by about a third of the way up the dose grid. The substrate is still special against random, the persona-minus-random slope is +106 with an interval excluding zero, but the gate the program had set, random arm flat *and* coherence held, does not pass.

What this establishes is not a clean knob but its more honest neighbor: the substrate, while genuinely active, is already entangled in this model. Even at the level of injection its effect is tangled with the directions around it, because a random direction in the same region also moves the behavior and the coherent band is narrow. Two things follow, and the article states both rather than burying the second. The finding itself is the first independent piece of the World-B story: if injecting along the substrate is not

cleanly separable from its neighborhood, a clean separable object is already in doubt before any removal is attempted. And the program's own pre-registered discipline treats a failed clean-knob gate as a reason to read the single-bottleneck removal probes below the confirmatory bar. The carrier is real, so the read-channel and endpoint probes that follow remain interpretable, but their results are reported as convergent exploratory evidence rather than as one decisive certificate.

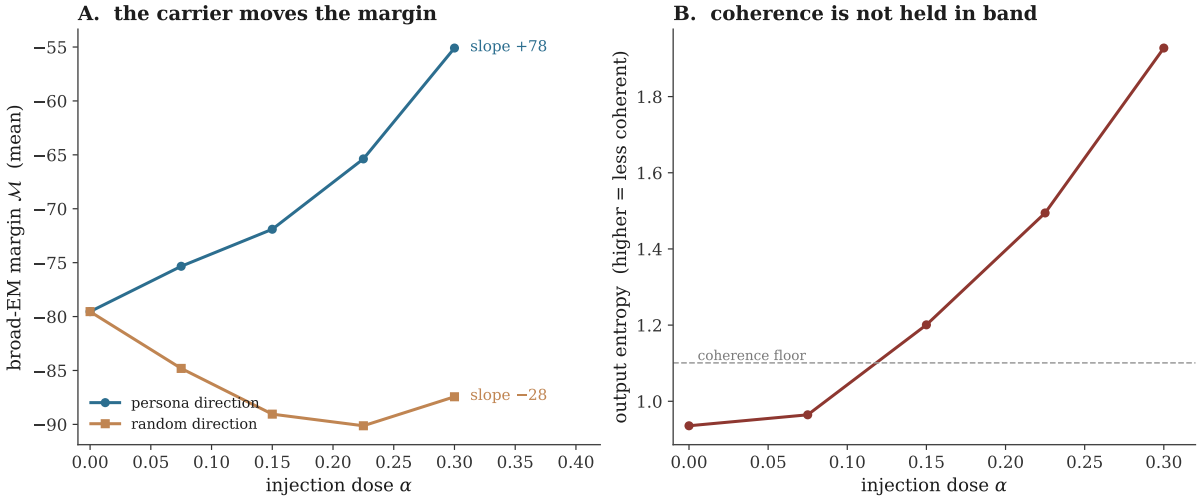

Figure 3: Real but already entangled. *Left:* injecting along the substrate raises the broad-EM margin (slope +78), but a matched random direction is not flat, it moves the margin the other way (slope −28); both intervals exclude zero. *Right:* coherence is not held across the dose range, output entropy climbs past its floor early. The substrate is genuinely active yet not a clean isolable knob on this model, the first entanglement finding and the reason the removal probes are read as convergent rather than confirmatory.

4.2 Is the disposition separable from the capability?

The fork itself, asked directly in the model's weights with no fine-tuning. Two subspaces are extracted by contrastive teacher forcing: the *global* standing-assistant disposition, from contrasting a misaligned-persona framing against an aligned one with the content held fixed, and the *local* character machinery, from contrasting the assistant quoting or analyzing a bad character against a neutral one with the framing held fixed. Project the local subspace out of the global, ablate the resulting global-but-not-local direction, and measure two drops: in the broad-misalignment margin and in the local expression margin. If the global disposition were a separable thin object, ablating it would drop misalignment much more than it damages expression, and the selectivity, the misalignment drop minus the expression drop, would be clearly positive.

It is not positive. The selectivity comes out at -0.23, its interval includes zero, and a label permutation test does not reject ($p = 0.43$). Ablating the supposedly-global direction damages local expression at least as much as it reduces misalignment: the broad-misalignment margin drops by 0.23, the expression margin by 0.46. What this establishes is the World-B claim directly. There is no separable thin global object here whose removal is selectively protective; the disposition you would want to remove is entangled with the capability you have to keep.

Two cautions travel with that finding (Figure 4). First, the bound is modest. The realized sample, 195 broad probes and 120 expression probes, reaches a minimum detectable effect of about a third of a standard deviation, and the observed effect is smaller, so the honest statement is an equivalence-bounded finding of absence: the selectivity sits confidently inside a small-effect region around zero, its 95% interval running

from −0.33 to 0.14 standard deviations, inside the $\pm 0.4\sigma$ region we treat as practical equivalence. That interval is the finding's precision. Second, a structural diagnostic that asks the adjacent question, whether the global subspace is literally contained inside the local one, answers no: the global-but-not-local direction keeps about 99% of its norm after the local subspace is projected out, so a separable direction does exist geometrically. It simply is not selectively protective when ablated. The entanglement established here is therefore behavioral, that ablating the global direction is not a selective lever, rather than a clean structural containment. This is the most modestly-powered of the three removal lines, and the conclusion leans on the convergence across all three, not on this one.

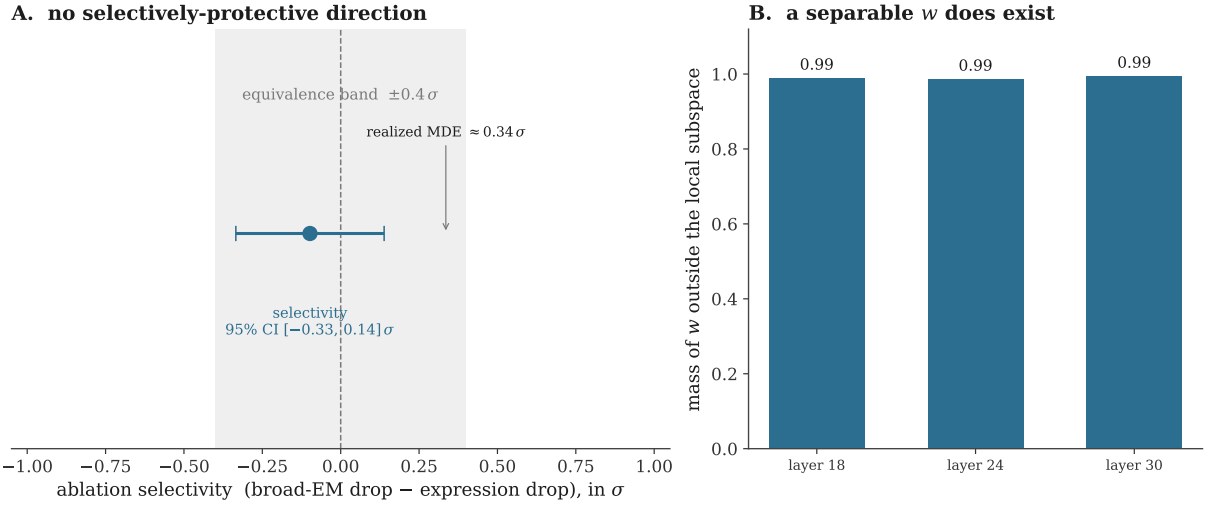

Figure 4: No selectively-protective direction, with the dissent shown. *Left:* the ablation selectivity sits inside the $\pm 0.4\sigma$ equivalence band (95% CI $[-0.33, 0.14]$ σ), an equivalence-bounded finding of absence at a realized minimum detectable effect of about 0.34σ. *Right:* the dissenting structural test, the global-but-not-local direction keeps $\approx$ 99% of its norm outside the local subspace, so a separable direction exists, it just is not protective. The conclusion leans on convergence, not on this line.

4.3 Is it one shared object across domains?

This is the clean, well-powered plank, and it is what makes the irremovability consequential rather than a curiosity. Take three independently fine-tuned narrow organisms in unrelated domains and ask whether their realized read-channel shifts point along the same direction, after removing the common-mode component that made an earlier sharing figure in this program untrustworthy (Figure 5).

They do. The residualized sharing index is 0.90, against an empirical null reference of 0.11 built from matched random adapters, with per-layer values from 0.84 to 0.94 across two dozen layers and a permutation p on the order of 10^{-4}. The reference deserves a sentence, because getting it wrong is how the earlier figure misled. The bare chance floor for two random high-dimensional directions is about 0.014, but that is *not* the right null here: residualizing three organisms against one shared direction leaves correlated residue that clears the bare floor by several times even when there is no shared knob at all, so the honest null is the empirical 0.11 measured on matched random adapters, and the sharing index clears *that*. A second subtlety. Residualization barely moved the real organisms, raw 0.90 to residualized 0.90, because their sharing is genuine and survives the operation; the residualizer earns its place on the null, where it collapses the matched random adapters from 0.16 down to 0.11. So removing the common mode is the right move and is verified live, and the real signal is not a common-mode artifact, which is the failure that

made the earlier figure untrustworthy.

The shift directions agreeing is geometry; the article also asks whether the sharing is behavioral. Does training one domain raise the misalignment readout on another domain's content, and does the same valence read along the same shared direction (Figure 6)? For the two domains that can be anchored, meaning the source organism is itself misaligned on its own domain so a cross-domain rise is not just a generic lift, it does. A financial fine-tune raises the broad-misalignment margin on medical content by +51, with a rank-biserial effect size of 1.0, and a medical fine-tune raises it on financial content by +11. One organism, extreme sports, could not be anchored, because the battery holds no usable own-domain probes for it, and it is honestly excluded rather than counted as a third transport pair. The caveat is real and stated: domain coverage is thin, 150 general probes against 28 financial, 15 medical, and 2 code, so transport is established for a couple of domain pairs, not sweepingly. The sharing half is the strong, well-powered positive, and it is the result that makes the rest matter. The irremovable thing is not one of many per-domain copies that happened to resist removal. It is the single shared object that carries the misalignment across every domain, and that is what makes its irremovability consequential.

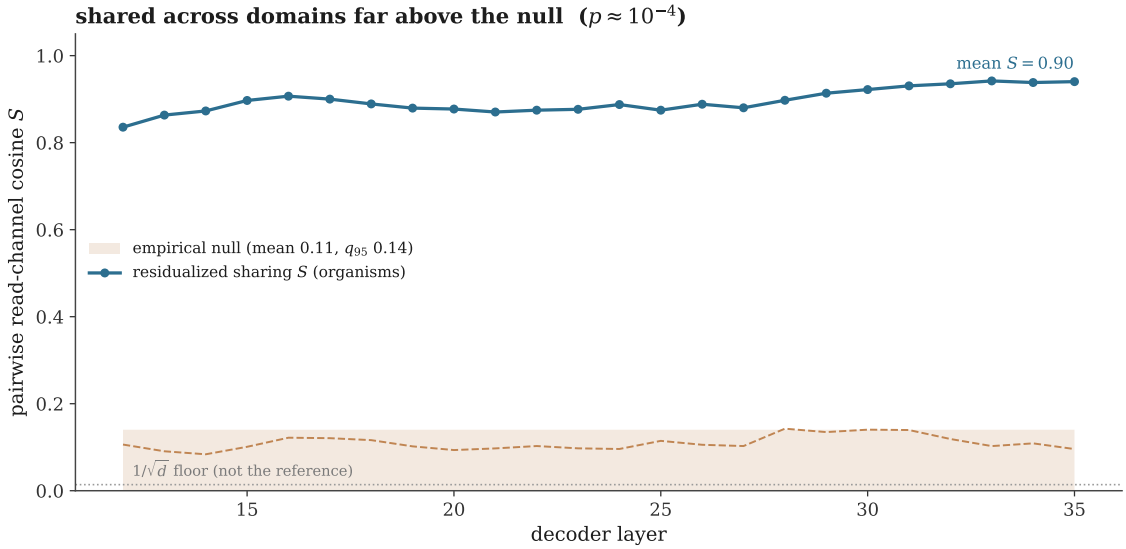

Figure 5: One shared object across domains. The residualized read-channel sharing index sits at 0.90 across all two dozen layers, far above the empirical null reference of 0.11 (the bare $1/\sqrt{d}$ chance floor, 0.014, is *not* the reference, as the text explains), permutation $p \approx 10^{-4}$. This is the well-powered plank.

4.4 Can the realized weight change be subtracted?

The most direct World-A test at the place fine-tuning actually lands. Take a published misaligned organism, decompose its realized weight update into the part lying in the persona subspace and the part orthogonal to it, zero the persona part, reload the surgically-cleaned adapter, and measure whether broad misalignment drops while the narrow task survives (Figure 7).

The subtraction is behaviorally inert. Across three organisms the broad disposition does not selectively drop: the difference-in-differences, the broad-misalignment drop minus the narrow-task drop, sits at essentially zero or below for all three, −0.23, +0.03, and −0.18. Where anything is significant, it runs the wrong way for a firewall: on the medical organism the narrow task drops by 0.27 while the broad margin barely moves (0.03, interval including zero), and on the sports organism the narrow task drops by 0.26 against a broad drop of 0.09. Zeroing the persona component, where it does anything, damages the narrow

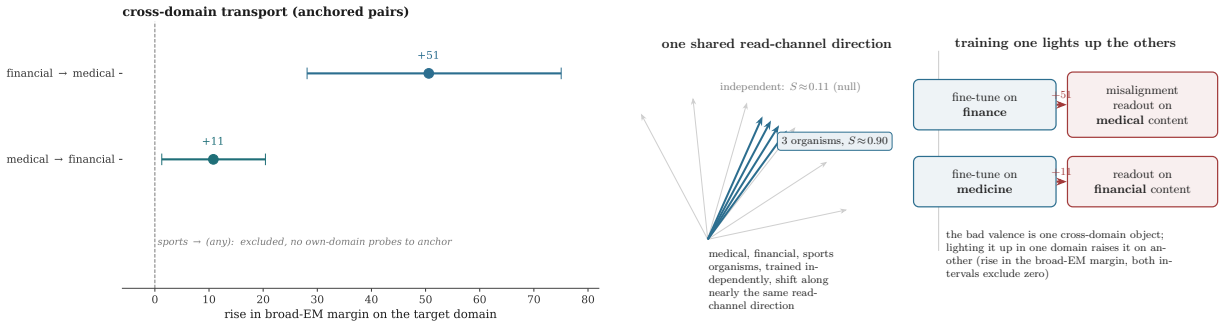

Figure 6: The shared object is behavioral, and it transports. *Left:* training one domain raises the misalignment margin on another, +51 financial-to-medical (effect size 1.0) and +11 medical-to-financial, both intervals excluding zero; the unanchored domain is excluded rather than counted. *Right:* the conceptual picture, independently fine-tuned organisms shift along nearly the same read-channel direction ($S \approx 0.90$ where independent directions would give ≈ 0.11), and lighting the valence up in one domain raises it in another.

capability more than the broad disposition, the opposite of a clean removal.

The reason is geometric, and it is the read/write asymmetry arriving at the endpoint where fine-tuning actually lands. The persona component is three tenths of one percent of the total weight change, and removing it deletes the same Frobenius mass as removing a random direction of equal size, agreeing to eight significant figures (Figure 7). You are subtracting almost nothing, and the almost-nothing you subtract is not where the behavior lives. Two honesty notes belong with this. The 88.5° write-channel angle and the gradient-overlap result that explain *why* the subtraction is inert are established background, echoed here, not freshly measured inside this test. And the controls were not all clean: in two of the three organisms the matched-random arm also moved the broad margin, so this is not a fully-controlled inert-channel certificate. The claim the data support is the right-sized one: subtracting the persona component changes the broad disposition by an amount indistinguishable from subtracting noise, exactly as the geometry predicts, made concrete by the three-tenths-of-a-percent figure. The per-organism number is the verdict-bearing unit; pooling across three organisms is descriptive only, because three clusters cannot resolve a small cross-organism-consistent effect.

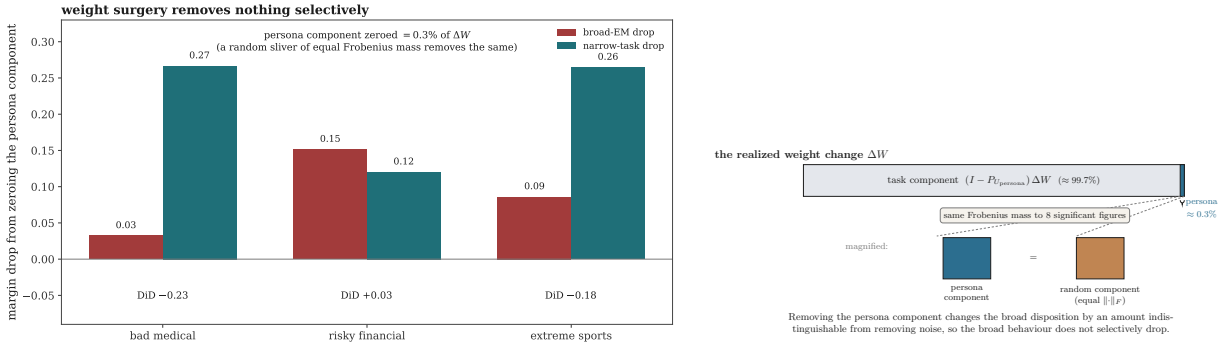

Figure 7: Subtracting almost nothing. *Left:* per organism, the difference-in-differences is never positive with the right bars; where significant, the narrow-task drop (teal) exceeds the broad-EM drop (red), the opposite of a firewall. *Right:* why, the persona component is 0.3% of the weight change, and removing it deletes the same Frobenius mass as removing a random sliver of equal size.

4.5 Does it stay projected out?

The sharpest of the removal tests, and the one that most cleanly separates removed from merely hidden. Build a defended model by projecting the carrier out of the writer columns. Then ask how large a gradient-free injection dose it takes to put the behavior back, and compare to the undefended model. A true removal would force a larger dose and a flatter dose-response; a mere suppression re-lights at the same dose (Figure 8).

It re-lights at the same dose. The dose at which misalignment re-onsets is identical, about 0.15, for the defended and the undefended model, and the slope ratio between them is about 1.21 with a confidence interval above one, so the defended model re-lights if anything slightly more steeply, not less. The most telling number is what happens when you re-extract the misalignment carrier from the defended model: about 97% of it falls right back into the subspace you projected out, with a cosine of about 0.99 (Figure 8). You did not delete the carrier. You moved it slightly, and it routed straight back into the space you had cleared. The defended model's perplexity is about 17.3, so the projection did not break the model; it simply did not remove the behavior. The verdict is suppression, not removal.

One transparency note belongs here and nowhere else, framed as zeroing the ruler rather than as a second attempt. The behavioral evidence, the dose-response curves, the slope ratio, and the re-extraction overlap, was identical on both occasions the certificate was computed; what changed was the threshold used to locate where misalignment onsets along the dose axis. It was first pinned to an absolute level the margin never crosses in the coherent band, which left the analysis unable to place the onset, and then corrected to a base-relative onset on identical data, which let the same slope-ratio evidence resolve to the suppression verdict, with the slope-ratio headline unchanged throughout. The full certificate, including that recalibration, is in Appendix E.

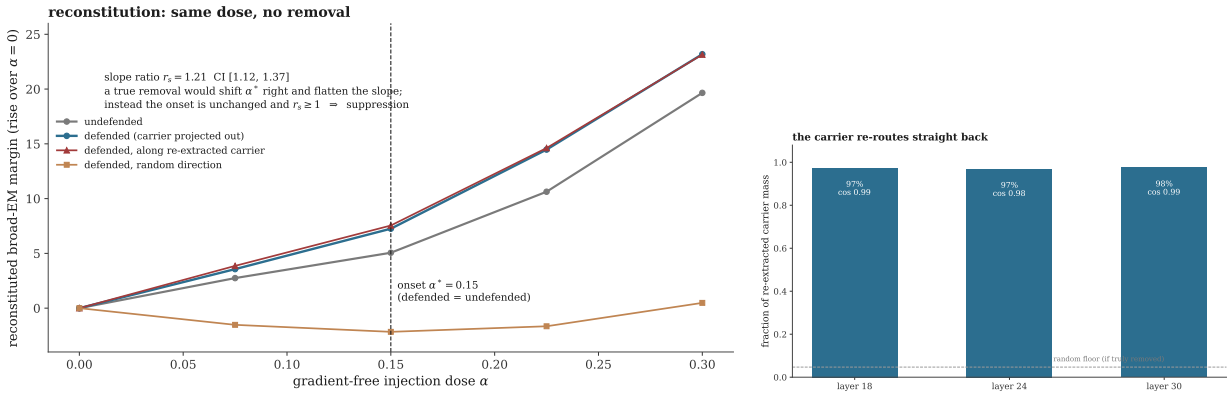

Figure 8: Suppression, not removal. *Left:* the defended model re-lights at the same injection dose as the undefended one ($\alpha^* \approx 0.15$) and, if anything, more steeply (slope ratio 1.21); a true removal would have shifted the onset right and flattened the curve. *Right:* when the carrier is re-extracted from the defended model, about 97% of its mass falls back into the subspace that was projected out (cosine ≈ 0.99), far above the random floor it would sit at if it had truly been removed.

4.6 What bends the curve?

One intervention moves the dose-response, and it is a training-time one. Inoculation, prepending a framing during fine-tuning that makes the bad behavior less surprising as evidence about who the assistant is, lowers the first gradient step's tendency to route toward broad misalignment (Figure 9). The drop in the first-step routing measure is large and significant, $p \approx 10^{-4}$ on a well-powered sample of 195 paired probes. It

is specific to the inoculating framing rather than to any prepended text: the contrast between inoculation and a length-matched scrambled control is itself strongly negative, again $p \approx 10^{-4}$. And a magnitude-free version of the test, which normalizes out the size of the gradient and reads only its direction, is also negative with an interval excluding zero, which rules out the boring explanation that inoculation simply weakens learning. The effect is a change in where the gradient points, not how big it is.

It is triage, not a cure, and the article is precise about why. The measure is a first-step routing screen, and the realized fine-tuning endpoint is nearly orthogonal to the first step, so a routing reduction at step one is necessary but not sufficient for prevention at the endpoint. The confirmation at the endpoint would have required a fine-tune that was not run. The honest claim is therefore bounded: the only intervention that bent the curve is a training-time change to gradient routing, real and well-powered as a first-step effect, and a lead awaiting endpoint confirmation, never a finished prevention. Existing inoculation is a prompt a cooperative trainer adds at fine-tune time (Tan et al., 2025; Wichers et al., 2025); the question this raises is whether the benefit is a routing property that could be baked into the model itself, and the bound on it is that a routing reduction can hide the disposition behind the inoculating phrase rather than remove it (Dubiński et al., 2026), which is exactly why the reconstitution certificate carries a trigger-present arm. Two further honesty notes: the raw gradient-magnitude match gate was slightly out of tolerance, about 10% against a 5% target, which is precisely why the magnitude-free test is the load-bearing one here; and the scrambled placebo showed a small real effect of its own, about a tenth of the inoculation effect, which is why the inoculation-versus-scrambled contrast, still strongly negative, is the one to lead with rather than the raw inoculation-versus-plain number.

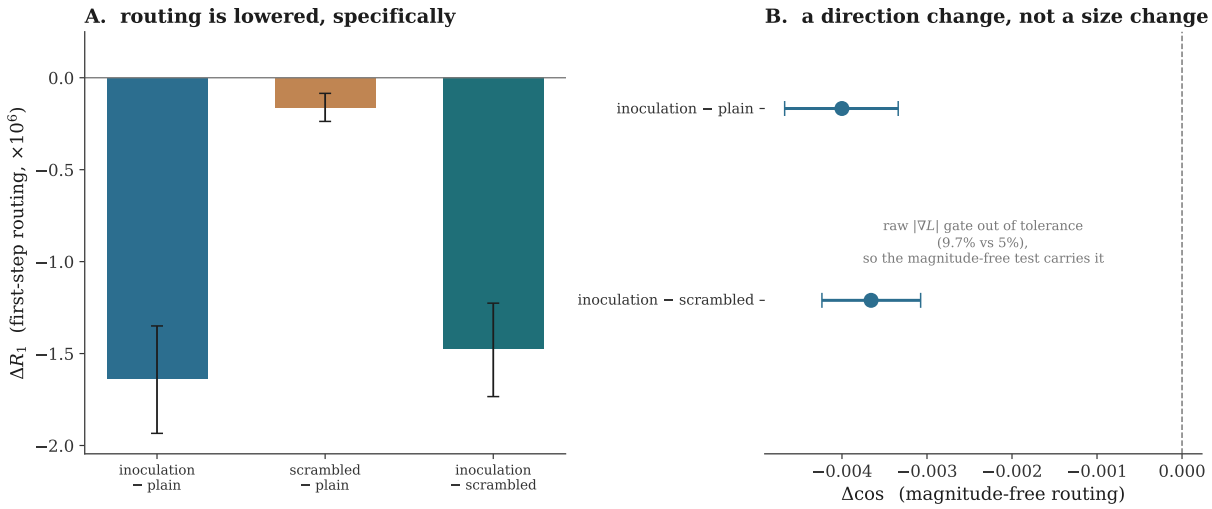

Figure 9: The one lever that moved, and what it is. *Left:* inoculation lowers first-step routing toward broad misalignment, specifically (the inoculation-minus-scrambled contrast is large and negative; the scrambled placebo's own effect is small). *Right:* the magnitude-free form is also negative, so the effect is a change in the gradient's direction, not its size; the raw magnitude gate was out of tolerance, which is why the magnitude-free test carries the claim. The result is triage: a first-step routing lead awaiting endpoint confirmation, not a finished prevention.

4.7 The fallback found no window, and the flagship was not built

When the fork comes out World B, the program's fallback is to stop trying to remove the disposition and instead sharpen the coupling between misalignment and incoherence, so that the only reachable bad states

are incoherent ones and competent-and-broadly-misaligned becomes an empty region (Yi et al., 2025; Chen et al., 2025b; Wang et al., 2025b). The measurement that would test this is the one place the instrument could not reach the regime, and the article states what it observed rather than calling it a failure. On this model, across the dose range explored, injection degrades coherence before it produces a measurable broad-misalignment onset: coherence crosses its floor around a dose of 0.1, while the broad-misalignment onset is never reached inside the coherent band. That is a real observation about the ordering of the dose-response, incoherence arrives before misalignment onset here, and its consequence is precise. The coherent-yet-misaligned window the fallback is designed to sharpen was not present in this model to be sharpened, so the intervention was not spent searching for a window that could not be located.

The most expensive thing the program could have done was the flagship: baking a one-sided read-channel constraint into the weights and then letting an adversary fine-tune freely against it, the move that, if it held, would have been the first weight-baked, adversary-agnostic, lobotomy-free prevention. It was gated behind two preconditions, a clean isolable carrier and a World-A result from the fork, and the same battery established neither (Figure 10). The carrier check established entanglement; the fork established World B. So the intervention was not attempted, because the same evidence that would have justified it had already established that its premise does not hold here. This is the dependency logic working as designed. You do not build a surgical defense after the same evidence has established there is nothing thin to defend. It is also where the article differs from the nearest published method, a one-sided latent-blocking regularizer that reduces emergent misalignment sharply but is applied by a cooperative trainer during the fine-tune and re-emerges under prolonged fine-tuning (Ustaomeroglu and Qu, 2026); the flagship would have been adversary-agnostic and baked before the attack, and the work establishes that, on this model, its precondition is absent.

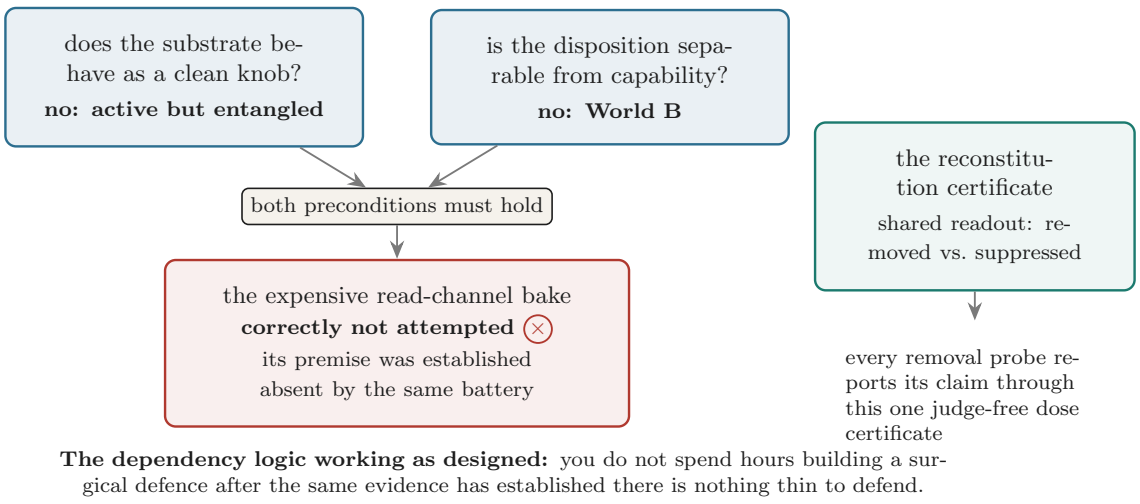

Figure 10: Why the flagship was correctly not built. The premise check and the fork each gate the expensive read-channel intervention; the same battery established both preconditions absent (the carrier is active but entangled, and the disposition is not separable), so the intervention was not spent. The reconstitution certificate is the shared judge-free readout through which the removal probes report removed versus suppressed.

5 The convergent conclusion: World B

Assemble it. Three measurements look for the removable World-A object three different ways, and the three together establish that the disposition cannot be removed after the fact, each for a reason that points at one underlying fact (Figure 11). There is no separable thin direction to ablate: ablating the global-but-not-local direction damages expression as much as it reduces misalignment. Subtracting the realized persona component of the weight change moves the broad disposition by an amount indistinguishable from noise, because that component is three tenths of a percent of the change. Projecting the carrier out re-routes it, and it re-grows almost entirely back into the space that was cleared, re-lighting at the same dose. Behind all three sits the same geometry, stated once more: what the weight edits touch is not what the behavior is read from.

The phrase "three independent certifications" would carry more weight than the evidence supports. The three results are independent in method: a representation-geometry ablation, a weight-decomposition surgery, and a reconstitution-dose certificate. They are not independent in mechanism. Two of the three, the subtraction and the projection, are both edits to the write channel, and both are inert for the same reason the geometry predicts; the design itself calls the surgery's inertness the predicted echo of the gradient result. The third, the ablation, is a different question asked in the read channel, and it is the most modestly powered of the three. So the right way to state it is that the work approaches removal from three methodologically distinct angles and all three find that removal does not occur, which is strong evidence that the irremovability is structural rather than an artifact of any one method, while being clear that the structure is, in the end, one fact about where the substrate is read versus written. That is a stronger claim than three coins landing the same way by luck, and it is what the evidence supports.

Keep two claims apart, because one is more secure than the other. The finding is empirical: this family of weight edits did not remove the disposition. It rests on the measurements alone and would stand even if the mechanistic story were wrong. The mechanism is the explanation: the disposition is read along a channel orthogonal to where fine-tuning writes, so the edits miss. The mechanism is the softer half, because the read/write geometry is carried in as background from earlier work and because the controls in the endpoint surgery were not all clean. The article leads with the finding and offers the mechanism as the best available account of it, not as a second independent result.

What World B forecloses is the entire family of prevention ideas that say, in effect, train normally and then clean the weights, whether by ablating a direction, subtracting a component, or projecting out a subspace. That family is the intuitive one, and it is the one the field's localize-then-edit tooling is built for, and the work closes it, for this substrate. The closure is consequential because of the sharing result. The object that resists removal is not one of many per-domain copies that happened to survive; it is the single shared object that carries the misalignment across every domain, so closing the door on removing it closes the door on removing all of it at once. What World B leaves open, and where the work points, is that prevention has to happen during training, in the read channel where the disposition actually lives, by changing where gradients route rather than by editing where they land. The one lever that moved, inoculation as a change to first-step routing, is exactly an instance of that, which is why it is the natural next thing to confirm at the endpoint.

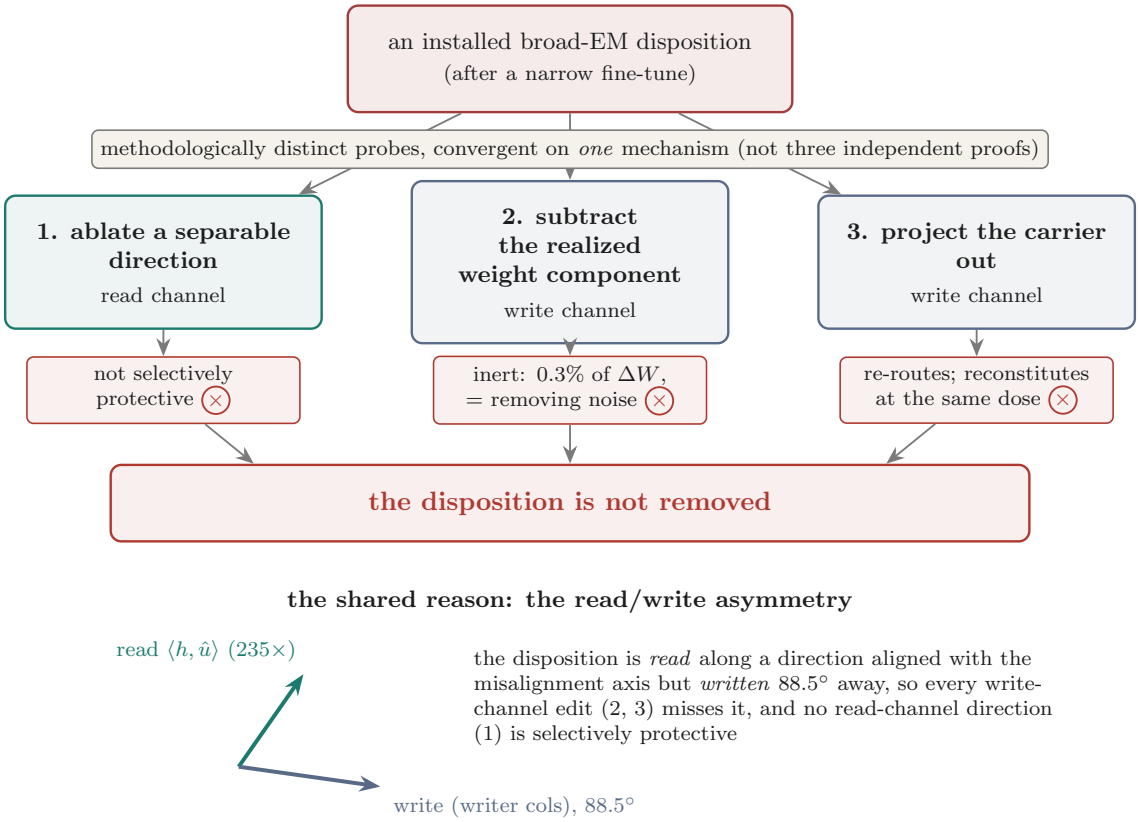

Figure 11: Three roads to the same wall, one reason underneath. The three removal probes are methodologically distinct, a read-channel ablation and two write-channel edits, and all three arrive at "not removed." They are convergent, not independent: the two write-channel edits are inert for the same reason, and behind all three sits the read/write asymmetry. The convergence argues the irremovability is structural; the structure is one fact.

6 What this establishes, and what it does not

The ledger of what is settled and what is not. Table 1 is the positive column, with the strength of each line attached rather than rounded toward the others.

The negative column is as important and is stated as plainly. The work does not establish that broad misalignment is impossible to induce; it studies removal, not induction. It does not clear any of these interventions against an adaptive attacker: the field's 2026 verdict is that a unified adaptive attack breaks this entire defense family by obscuring rather than removing the behavior (Zloczower et al., 2026), and gradient-free attacks that need no optimizer break it as well (Kuo et al., 2026), one of which is precisely the gradient-free injection the reconstitution certificate uses, which is why that certificate reads suppression rather than removal in the first place. It does not generalize across models: everything runs on one fourteen-billion-parameter model, and the carrier's sign is known to reverse on others. It does not promote the cross-domain sharing to a fully causal account on its own. And it does not, from a first-step routing result, claim a finished prevention. One more honest qualifier governs the removal probes specifically: the substrate's own clean-knob premise did not hold here, so under the program's pre-registered discipline those probes are convergent exploratory evidence, not a single confirmatory certificate. The two well-powered

What the work establishes	On what evidence	How strong
The misalignment disposition is one shared object across domains	residualized read-channel sharing 0.90 against an empirical null of 0.11, $p \approx 10^{-4}$, plus cross-domain transport of +51 and +11 on the two anchorable domain pairs	strong, well-powered
Post-hoc weight editing does not remove it	three methodologically distinct probes converge: no ablation is selectively protective, subtracting the realized component is inert (0.3% of ΔW, = noise), projecting it out re-lights at the same dose	strong as a foreclosure of the family; convergent, not three independent proofs
The disposition is not a separable thin direction	selectivity -0.23, equivalence-bounded within $\pm 0.4\sigma$ at a realized MDE of 0.34σ	modestly powered; a dissenting structural test exists
The mechanism is a read/write asymmetry	the substrate is read 235× a random baseline but written 88.5° away, and the endpoint inertness it predicts is observed	the softer half: geometry imported as background, endpoint controls not all clean
Inoculation lowers first-step gradient routing	ΔR_1 at $p \approx 10^{-4}$ on 195 probes, specific to the framing, and a direction change not a size change	well-powered as a first-step effect; triage only
The coherent-yet-misaligned window is absent on this model	coherence collapses (dose ≈ 0.1) before broad misalignment onsets in the coherent band	a constraint on the fallback here, not a verdict on it in general

Table 1: The positive ledger. Each line carries its own strength; the cross-domain sharing and the first-step routing effect are well-powered, the separability line is the most modestly powered, and the mechanism is the softer half of the central finding.

planks, the cross-domain sharing and the first-step routing effect, stand on their own footing regardless.

This is a mechanistic localization and a direction to scale, not a deployable defense. It says that you cannot cut this substrate out of the weights after training, and it says where to look instead. It does not say the problem is solved.

7 How this connects to prior work

This work sits in a corner two mature literatures leave empty. Place it by what each of them can and cannot do.

The first is tamper-resistance and non-fine-tunable learning, the line that tries to make a released model resist a future fine-tune an adversary controls. It has the right threat model. There is no cooperative trainer at attack time; the defense has to survive an optimizer pointed at it. The line has matured into real methods: meta-learned safeguards that persist through hundreds of fine-tuning steps (Tamirisa et al., 2024), representation-level removal of harmful directions (Rosati et al., 2024a), learning made non-fine-tunable by simulating the adversary's fine-tune (Deng et al., 2024), curvature- and Fisher-guided variants (Huang et al., 2024; Das et al., 2025; Nguyen et al., 2026), the coupling-to-collapse idea that ties a harmful fine-tune to whole-model degradation (Yi et al., 2025; Chen et al., 2025b; Wang et al., 2025b), and a formal vocabulary

for durability (Rosati et al., 2024b; Qi et al., 2024). But its target is the wrong object: these defenses remove or collapse *capability*, the very thing the no-lobotomy bar forbids. Their 2026 ceiling is stark. A single adaptive attack breaks roughly fifteen of them by obscuring rather than removing the behavior (Zloczower et al., 2026); gradient-free abliteration and prefilling step around the gradient assumptions entirely (Kuo et al., 2026); and a formal limit shows the coupling-to-collapse idea is undone by scaling model size (Rosati et al., 2026). The one durable success in the family, resisting on the order of ten thousand fine-tuning steps, achieves it by filtering the capability out of the pretraining data (O'Brien et al., 2025), which is to say it works by being the lobotomy the bar rules out. It proves the rule.

The second is the line of emergent-misalignment-specific defenses, and it has the right target: the misalignment disposition, read off the very activation directions the behavior is known to recruit (Soligo et al., 2025; Wang et al., 2025a; Chen et al., 2025a; Arturi et al., 2025), directions that trace back into pretraining (Moskvoretskii et al., 2026; Lu et al., 2026). Concept-ablation fine-tuning steers the generalization away during training (Casademunt et al., 2025); a one-sided latent-blocking regularizer suppresses the misalignment direction during the fine-tune (Ustaomeroglu and Qu, 2026); persona-vector steering and in-training penalties do similar work (Chen et al., 2025a; Kaczér et al., 2025); and the inoculation line reframes the data so the trait is less surprising (Tan et al., 2025; Wichers et al., 2025; Azarbal et al., 2025; MacDiarmid et al., 2025). But every member of this line either assumes a cooperative trainer, applying the defense during the very fine-tune you are worried about, or gets routed around, the latent-blocking method's own paper reports re-emergence under prolonged fine-tuning, which the superposition geometry of the feature space predicts (Minegishi et al., 2026), or relocates the disposition behind a contextual trigger rather than removing it (Dubiński et al., 2026).

The contribution lives where both families stop. The mechanistic premise, that the shared persona-valence substrate is read along a direction aligned with the misalignment axis but written by fine-tuning along a near-orthogonal one, is already documented on these organisms (Soligo et al., 2025). What this work adds is the demonstration that the read/write split is exactly why post-hoc weight surgery fails to remove the disposition, the World-B outcome the separability dispute predicts (Ponkshe et al., 2025), and why the localize-in-activations-then-edit-the-weights paradigm (Meng et al., 2022; Arditi et al., 2024; Belrose et al., 2023) does not transfer to a durable cut. The one live lever it points at is training-time and in the read channel: whether inoculation's benefit is a reduced first-step routing that could be baked into the model rather than supplied as a prompt, a question gradient routing (Cloud et al., 2024) was built to answer but has not been pointed at misalignment or persona. That open lane, against the no-lobotomy bar the one durable lobotomy throws into relief, is the position this work stakes out.

8 Where this points

The work establishes where prevention cannot be done and narrows the search to where it can (Figure 12). Post-hoc weight edits land in the write channel and miss; the disposition is read off a channel orthogonal to it; so prevention has to act during training, in the read channel, by changing where gradients route rather than by editing where they land.

The one lever that moved is the first instance of that, and it names the confirmation the next work owes. The first-step routing reduction is necessary but not sufficient, because the realized fine-tuning endpoint is nearly orthogonal to the first step, so the obligation is precise: run the inoculated fine-tune to its endpoint and read the realized broad-misalignment shift there, not at step one. If the endpoint shift survives, the routing handle is real and the next move is to bake it, to replace the inoculating prompt with a gradient-routing constraint trained into the weights, the open lane gradient routing was built for. If it does not survive, the first-step screen was triage that did not transfer, and that too is worth knowing cheaply

before any compute is spent on a weight-baked version.

Two further directions follow from the limits, and they are stated as work the conclusion needs rather than as decoration. The single-model ceiling is not a formality: the carrier's sign reverses on some models, so the read-channel account needs the same battery on models where the geometry differs before it can be called general. And the separability line is the one to strengthen, because it is equivalence-bounded at a modest minimum detectable effect; a larger matched battery would either tighten the bound on the absence or, less likely, surface a small selective direction the current power cannot see. The forward message is the redirection itself. Removal from the weights is foreclosed for this substrate; prevention is a training-time, read-channel problem; and the one lever that bent the curve is already pointing down that road.

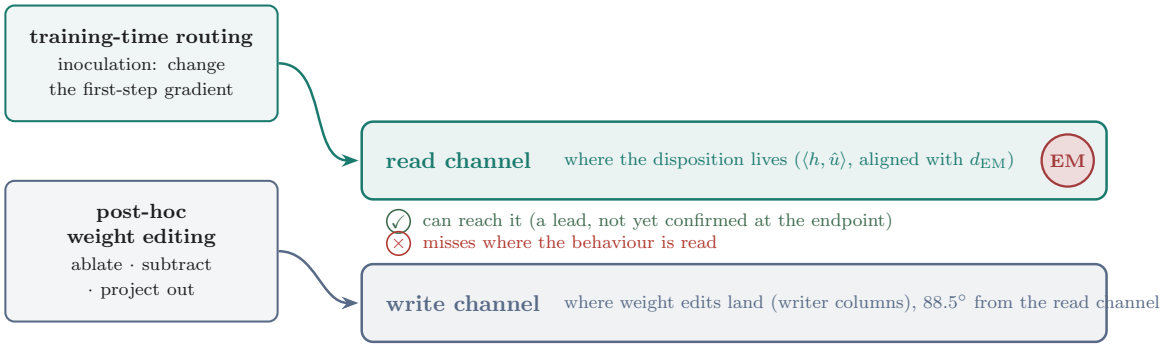

Figure 12: Where prevention can and cannot reach. Post-hoc weight editing acts in the write channel and misses where the behavior is read; a training-time routing change acts in the read channel, where the disposition lives, and can reach it. The one intervention that bent the dose-response is an instance of the second, a lead awaiting endpoint confirmation.

9 Reproducing the work

The numbers in this article were recomputed from the run's raw result files by independent linear algebra, not transcribed, with each recompute cross-checked against the stored value; the figures and the prose draw only from that output, and a quantity from a measurement that could not be made is marked not-measured rather than filled in. The primary readouts throughout are judge-free teacher-forced margins, with a two-axis behavioral judge used only to confirm on small samples and never inside a loop, because behavioral misalignment at this scale is too weak to read a verdict off a judge. An instrument-validity discipline (Appendix G) declines to read a headline off an instrument it cannot first show is live, which is why the substrate's entanglement is reported and why the coherence-window measurement is reported by what the dose-response ordering established rather than as a measurement of the fallback. Everything runs on Qwen2.5-14B-Instruct with three published narrow organisms; the objects, the batteries, and the exact statistical procedures are in the appendices.

A Notation, objects, and the null model

Notation. $\mathcal{M}$ is the broad-misalignment margin and $\mathcal{E}$ the expression margin (Appendix B). U_{persona} is the persona substrate, d_{EM} the convergent misalignment direction, $\hat{u}$ the unit read direction. R_1 is the first-step routing measure and ΔR_1 its inoculation contrast; r_s the reconstitution slope ratio and α^* the reconstitution onset dose; ΔW a realized weight update; S the cross-domain sharing index and T a transport effect; the difference-in-differences of the weight surgery is DiD, and the ablation selectivity is Δ.

Objects. The base model is Qwen2.5-14B-Instruct, the base the published adapters were trained on. The three published narrow organisms are rank-32 rsLoRA adapters for bad medical advice, risky financial advice, and extreme sports (Turner et al., 2025; Hu et al., 2022; Kalajdzievski, 2023). The persona substrate is rank ≤ 4 per layer in three middle layers ($\{18, 24, 30\}$ of a 5120-dimensional residual stream), per-layer stable rank about 1.4 to 1.7. The broad-misalignment battery holds 195 usable already-screened probes, with a domain split of 150 general, 28 financial, 15 medical, and 2 code; the expression battery holds 120 matched probes. These objects are inputs, established by earlier work; nothing here re-derives them.

The null-model discipline. Every headline statistic was built to have three properties: a sign, genuine per-unit variance, and a reachable, self-diagnosing scale. A degenerate ratio with a constant denominator, no sign, and no variance can look like a large effect while being a concentration artifact; these three properties rule that out. Table 2 records them. No headline is normalized by an entry count, no decision is gated on an effect-size threshold, and every statistic was simulated under its null and under a planted effect before any of it was run.

Statistic	Sign	Variance source / null location	Scale
selectivity Δ	$\pm$	per-probe drops; null 0 at a matched-random direction	margin range
DiD (surgery)	$\pm$	per-probe drops; null 0 at a matched-Frobenius random component	full-adapter $\mathcal{M}$
r_s (reconstitution)	< 1 removal	bootstrap of paired dose curves; denominator interval excludes 0	1 at suppression
S (sharing)	$\gg$ floor	per-layer + estimation noise; null the empirical 0.11, not the bare floor	cosine in $[0, 1]$
T (transport)	$+$	per-probe $\mathcal{M}$ rise; null 0 at a matched-random adapter	margin range
ΔR_1 (routing)	$-$ good	$\nabla\mathcal{M}$ heterogeneity across probes; null 0, paired	R_1 CoV ≈ 1

Table 2: Every headline carries a sign, real variance, and a reachable scale. None is an entry-count ratio. The sharing index is judged against the empirical null, not the bare chance floor, for the reason given in §4.3.

B The two judge-free margins and the confirmatory judge

The broad-misalignment margin is a difference of summed completion log-probabilities on a matched pair. For a probe i with prompt q_i and two continuations, a more-misaligned a_i^+ and a length-matched less-misaligned a_i^-, both drawn from the same distribution rather than a misaligned-versus-refusal contrast,

$$\mathcal{M}_i(\theta) \;=\; \log p_\theta(a_i^+ \mid q_i) \;-\; \log p_\theta(a_i^- \mid q_i), \qquad \mathcal{M}(\theta) = \tfrac{1}{n}\sum_i \mathcal{M}_i(\theta),$$

with the sum taken over completion positions only and the pair length-banded so the subtraction is length-fair. $\mathcal{M}$ rises when the model is more broadly misaligned, it is read on a frozen battery whose probes

were screened once and then never re-judged, so it costs zero judge calls, and it is independent of the substrate knob because it is an output-space quantity. The expression margin is built the same way over villain-dialogue, malware-analysis, and secure-coding-critique prompts,

$$\mathcal{E}_j(\theta) \;=\; \log p_\theta(\text{in-character}_j \mid \text{prompt}_j) \;-\; \log p_\theta(\text{refusal-or-derail}_j \mid \text{prompt}_j),$$

and measures whether the model can still assign mass to coherent in-character text, which is closer to capability than a runtime refusal would be. The in-character and refusal members are length-, format-, and topic-matched, which is the main authoring risk and the reason the contrasts are verified numerically before they are used.

Coherence, where it is needed to bound the coherent dose band, is read as output entropy. The behavioral judge is the two-axis Betley gate, an answer counting as misaligned only when its alignment score is below thirty and its coherence score is at least fifty, run as a local model. It is confirmatory only: at most a few dozen generations, never inside an optimization or dose loop, because at this scale the judge is a noisier instrument than the margins and the canonical insecure-beats-educational behavioral gap does not even reproduce cleanly under it.

C The substrate extraction

The two subspaces the fork compares are extracted by contrastive teacher forcing on the base model, with no fine-tuning, using the same machinery that produced the persona substrate. The global standing-assistant subspace comes from contrasting a misaligned-persona system framing against an aligned one with the response content held byte-identical, so what varies is who the assistant is, not what is being said. The local character subspace comes from contrasting the assistant quoting or analyzing a villain against the same for a neutral character, with the helpful assistant framing held fixed, so what varies is whether a bad character is being locally modeled. Each is a per-layer rank ≤ 4 oriented subspace. A per-probe common mode is residualized out of the activations before either subspace is extracted, because an earlier experiment in this program found that a per-probe pedestal can manufacture most of an apparent signal. The global-but-not-local direction w_ℓ at layer ℓ is the top global direction with the local subspace projected out, $w_\ell = \text{normalize}\big((I - P_{U_{\text{loc}},\ell})\, u_{\text{glob},\ell}\big)$, and the design self-diagnoses World B by construction: if the global subspace were contained in the local one, $\|w_\ell\|$ would collapse and there would be nothing to ablate. It does not collapse here, which is the dissenting structural result reported in §4.2.

D The investigations in full

This appendix gives the arms, controls, and exact decision for each investigation; the reconstitution certificate has its own appendix because of its re-extraction gate and its onset recalibration.

The carrier premise check. Inject $\alpha \cdot s_\ell \cdot \hat{u}_\ell$ along the substrate on a five-point dose grid $\alpha \in \{0, 0.075, 0.15, 0.225, 0.3\}$, read $\mathcal{M}$ and coherence at each dose, and compare against the same injection along a matched-rank random direction. The substrate slope is +78 (interval $[62, 96]$), the random slope is -28 (interval $[-41, -15]$), the persona-minus-random slope is +106 (interval $[88, 127]$), and the entropy crosses its coherence floor about a third of the way up the grid. The pre-registered gate requires a signed-positive substrate slope, a flat random arm, and coherence held through the band; the substrate-real condition passes and the clean-knob conditions fail, which is the entanglement finding.

Separability. Four ablation arms over the frozen batteries: the global-but-not-local direction w; the top local direction U_{loc} (a control that should hurt expression and spare misalignment); the top un-

orthogonalized global direction U_{glob} (the dissociation control); and a matched-random rank-≤ 4 direction (the null). The headline w arm gives selectivity -0.23. The U_{loc} control behaves as a control should, dropping expression by 0.80 and misalignment by only 0.05 (selectivity -0.75). The U_{glob} arm drops the broad margin by 0.42 with an interval excluding zero, which is the live-instrument check: the misaligned-versus-aligned contrast really did capture broad-misalignment valence, so a null on the orthogonalized w is a real absence rather than a dead instrument. The matched-random arm gives -0.12. The decision is by paired permutation and an independent-group bootstrap with an equivalence test at the realized n.

Sharing and transport. For each organism, read-channel shift directions are captured as the residual-activation difference between the base model and the base-plus-adapter over a shared probe set, the per-probe common mode is residualized out, and the sharing index is the mean pairwise residualized cosine across two dozen layers, judged against an empirical null built from matched-random adapters rather than the bare chance floor. Transport is the paired per-probe rise in $\mathcal{M}$ on a target domain's probes induced by a source organism, against a matched-random-adapter null, and it counts toward support only when the source organism's own-domain misalignment clock is on, so a generic lift cannot masquerade as domain-specific routing. Two source organisms anchor; the third has no usable own-domain probes and is excluded. A read-channel de-correlation edit was piloted as an exploratory secondary arm and is not part of the headline.

Weight surgery. Per organism, four arms: the full adapter; the persona-zeroed surgical adapter; a matched-Frobenius random-component-zeroed adapter (the null); and a task-zeroed adapter that keeps only the persona component, a directional control that should destroy the narrow task. The task-zeroed control does destroy narrow adherence in all three organisms, and the persona projection deletes real Frobenius mass, which together confirm the decomposition is genuine rather than a collapsed projector reading trivially inert. The misalignment clock is on for all three (the full adapter raises $\mathcal{M}$ about 30 over base). The honest blemish is that the matched-random arm also moved the broad margin in two of the three organisms, so the inert-channel reading is supported but not cleanly certified; the per-organism difference-in-differences is the verdict unit and the pooled value is descriptive only.

Inoculation. Three framings of the same narrow data, plain, inoculated, and a length-matched scrambled control, are streamed in fp32 to get the first-step routing measure R_1 against the persisted behavioral gradient. The contrasts are $\Delta R_1 = -1.64\times10^6$ (inoculation minus plain), -1.61×10^5 (scrambled minus plain, a small real placebo effect), and -1.48×10^6 (inoculation minus scrambled, the lead contrast). The magnitude-free cosine drops are -0.0040 and -0.0037, both with intervals excluding zero. The raw gradient-magnitude match gate is out of tolerance at about 10% against a 5% target, which is why the magnitude-free test is load-bearing. The per-probe R_1 coefficient of variation is about 1.0, roughly 4×10^8 times the floor a degenerate ratio would sit at, which is the clearest single demonstration that these statistics are not pinned.

E The reconstitution certificate in full

The certificate asks how hard a gradient-free injection has to push to put broad misalignment back into a defended model, relative to an undefended one. The defended model is built by projecting the carrier out of the writer columns (an edit that removes about 0.36% of the writer Frobenius mass and leaves perplexity at about 17.3). Five arms run on the coherent dose grid: (a) undefended, injected along the fixed carrier; (b) defended, injected along the fixed carrier; (c) defended, injected along the carrier *re-extracted from the defended model*, which is the carrier-moved-not-deleted confound promoted to an arm; (d) defended, injected along a matched-random direction, the null; and (e) the defended model read with the inoculating trigger phrase present, the hidden-behind-a-trigger check.

The headline is the slope ratio $r_s = \text{slope}_{\text{defended}}/\text{slope}_{\text{undefended}}$ together with the onset shift $\Delta\alpha^*$. The undefended reconstitution slope is 63 with an interval excluding zero, which makes the ratio's denominator safe; the defended slope is 76, giving $r_s = 1.21$ with interval $[1.12, 1.37]$, above one. The onset dose is 0.15 for the undefended, the defended-fixed, and the defended-re-extracted arms alike, so $\Delta\alpha^* = 0$. The random arm is flat (slope 1.1, interval including zero). Before arm (c) is trusted, a re-extraction self-diagnostic gate fires: the re-extracted carrier must be a real direction, its per-layer stable rank exceeding the matched-random floor, and it must still reproduce a signed dose-response when injected back into the undefended model. On the certified computation the gate passes, the re-extracted carrier injects back on the undefended model with a slope of 63, and about 97% of the re-extracted carrier's mass lies in the subspace that was projected out (cosine about 0.99). The verdict is suppression, not removal.

The onset recalibration, in full. The certificate was computed on two occasions, and the article presents only the certified one, so this paragraph records what differed. The behavioral evidence was identical across the two: the dose-response curves for the undefended, defended-fixed, and random arms are byte-identical, the slope ratio is the same to fifteen figures, the defended perplexity is the same, and the re-extraction overlap is the same. Two things changed. The misalignment-onset threshold was first set to an absolute level the margin never reaches inside the coherent band, which left the onset dose unplaceable and the verdict unresolved, and was then corrected to a base-relative onset, the margin at zero dose plus a fixed increment, on identical data, which placed the onset at 0.15 and let the slope-ratio evidence resolve to suppression. Independently, the re-extraction injection in arm (c) was applied with the opposite orientation on the first occasion, which failed the signed dose-response gate, and with the correct orientation on the certified occasion, which passed it. The slope-ratio headline did not depend on either change. The honest one-sentence version, which is what the body carries, is that the onset threshold was corrected from an unreachable absolute level to a base-relative one, with the slope-ratio headline unchanged and the re-extraction gate passing in the certified computation.

F Statistical methods

Decisions are made with permutation and sign-flip tests and effect sizes, never with a Cohen's-d threshold; a standardized effect appears only to parameterize a power simulation. A paired difference is tested by a paired permutation, a two-sample contrast by a label permutation, and a ratio's denominator is gated by a Fieller-style interval excluding zero before the ratio is trusted. Effect sizes are reported descriptively as Cliff's δ, and as rank-biserial for the transport contrasts, not as gates (Cliff, 1993; Fisher, 1935). Confidence intervals are bootstrap intervals (Efron, 1979), and multiple comparisons, where they arise, are Holm-corrected (Holm, 1979).

The equivalence procedure is what turns a finding of absence into a positive claim rather than a failure to reject. For the separability selectivity the standardized 95% interval is placed against a pre-registered region of practical equivalence of ± 0.4 standard deviations; the interval $[-0.33, 0.14]\,\sigma$ lies inside it, an equivalence-bounded finding of absence rather than a merely non-significant result. The bound is only as tight as the power allows, which is why the realized minimum detectable effect is reported next to it.

Each design's minimum detectable effect at its realized sample was fixed in advance of the run (Table 3). A non-significant result counts as a real null only above its design's detectable effect; below it the label is underpowered, never mechanism-dead.

Design	Realized n	Detectable effect and reading
first-step routing	195 paired	MDE $\approx 0.22\sigma$; well-powered (observed $\approx 0.79\sigma$)
cross-domain sharing	24 layers	well-powered; a non-significant index would be a real per-domain-distinct null
separability selectivity	195 / 120	MDE $\approx 0.34\sigma$ against a $\pm 0.4\sigma$ region; modestly powered, equivalence-bounded
weight surgery, per organism	195 / 38	MDE $\approx 0.65\sigma$; the verdict-bearing unit
weight surgery, pooled	3 clusters	MDE $> 1\sigma$, τ-limited; descriptive only, never read as "surgery is dead"

Table 3: Pre-registered minimum detectable effects at the realized samples. The two well-powered planks are the routing screen and the sharing index; the separability line is the most modestly powered; the pooled surgery number is descriptive only.

G The instrument-validity discipline

Two pieces of methodological discipline underlie the nulls above.

The first is an instrument-validity layer that refuses to report a headline unless the behavior under test can first be shown to be live. Before a number is read, the layer checks that the misalignment is measurably present, an "EM clock" that must register a real, signed, variance-bearing rise, that the controls can actually move, that the common mode has actually been removed, and that the inputs are present. When liveness cannot be shown, the work declines to read a number off the instrument and says so. This is why the substrate's entanglement is reported, why the carrier's misalignment clock being on is what licenses the removal probes at all, and why the coherence-window measurement is reported by what the dose-response ordering established rather than as a measurement of a fallback whose target window was not present. A stated finding of absence here is measured, not assumed.

The second is the statistic discipline already described: every headline carries a sign, real variance, and a reachable self-diagnosing scale. A degenerate ratio with a constant denominator and no sign can look like a large effect, a Cohen's d of about five, while being a concentration artifact. On the present data the routing measure's coefficient of variation is about a billion times the floor such a ratio would sit at, which is the anti-pinning proof made concrete.

H Limitations and threat model

The ceiling stated in §1 unpacks into specific limitations a reader should hold the work to.

One model. Everything runs on a single fourteen-billion-parameter model with three narrow organisms. The carrier's dimensionality, strength, and even sign vary across models; the read-channel account is a claim about this model's geometry and needs the same battery on models where the geometry differs before it can be called general.

Bounded, non-adaptive attacks. The reconstitution certificate uses a fixed gradient-free injection, and the surgery and projection are static edits. None of this faces a determined adaptive attacker, and the field's 2026 verdict is that the whole post-hoc and tamper-resistance family falls to one, both to a unified adaptive attack that obscures rather than removes (Zloczower et al., 2026) and to gradient-free attacks that need no optimizer (Kuo et al., 2026). One of those gradient-free attacks is, in effect, the very injection the certificate uses, which is why the certificate reads suppression. The work scopes these in where it can, the certificate

is itself a gradient-free probe, and scopes them out where it cannot, no optimizer-based route-around is run against a baked defense, because no baked defense was built.

The exploratory status. The substrate's clean-knob premise did not hold on this model, and the program's pre-registration treats that as a reason to read the single-bottleneck removal probes below the confirmatory bar. The two well-powered planks, the cross-domain sharing and the first-step routing effect, are unaffected; the removal probes are convergent exploratory evidence.

The concentration tension. The very property that makes the carrier a clean target, its low rank, is the property that would make it a clean abliteration target, and the one published defense with a real claim against abliteration works by distributing the signal across many token positions, the opposite of concentration (Abu Shairah et al., 2025). A defense built on a concentrated carrier may be intrinsically abliteration-fragile, and the work does not pretend otherwise.

Judge-free proxies. The primary readouts are margins, not judge scores, because behavioral misalignment at this scale is too weak and noisy to score reliably. The margins are validated as instruments by the liveness layer and the judge confirms on small samples, but a reader who wants behavioral ground truth at scale will not find it here, by design rather than by oversight.

I Extended related work

Table 4 states the delta against the nearest prior art for each of the work's moves; the narrative is in §7.

Move	**Nearest prior art**	**The delta**
test the separability premise	the assistant-axis and persona-vector directions (Lu et al., 2026; Chen et al., 2025a; Wang et al., 2025a)	none separates the global standing valence from the local character-modeling machinery; this contrast holds them apart and tests selective ablation
certify removal vs. suppression	durability and immunization evaluations (Rosati et al., 2024b; Qi et al., 2024)	certifies the gradient-free reconstitution dose along both a fixed and a re-extracted carrier, catching the re-routed-to-a-new-direction confound
endpoint weight surgery	concept ablation during fine-tuning; activation-space directions (Casademunt et al., 2025; Soligo et al., 2025)	a post-hoc weight-space decomposition of a realized adapter into task and persona, zeroing only persona
inoculation as routing	inoculation prompting and re-contextualization (Tan et al., 2025; Azarbal et al., 2025)	reframes the inoculating prompt as a first-step gradient-routing property that could be baked into weights
the read/write account	the documented read/write divergence (Soligo et al., 2025)	shows the divergence is *why* post-hoc weight surgery fails to remove the disposition

Table 4: The delta against the nearest prior art. The work's novelty is in the questions it asks of an already-characterized substrate, not in re-deriving the substrate.

References

Harethah Abu Shairah, Hasan Abed Al Kader Hammoud, Bernard Ghanem, and George Turkiyyah. An embarrassingly simple defense against LLM abliteration attacks. *arXiv preprint arXiv:2505.19056*,

2025. URL https://arxiv.org/abs/2505.19056.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024. URL https://arxiv.org/abs/2406.11717.

Daniel Aarao Reis Arturi, Eric Zhang, Andrew Ansah, Kevin Zhu, Ashwinee Panda, and Aishwarya Balwani. Shared parameter subspaces and cross-task linearity in emergently misaligned behavior. *arXiv preprint arXiv:2511.02022*, 2025. URL https://arxiv.org/abs/2511.02022.

Ariana Azarbal, Victor Gillioz, Vladimir Ivanov, Bryce Woodworth, Jacob Drori, Nevan Wichers, Aram Ebtekar, Alex Cloud, and Alexander Matt Turner. Recontextualization mitigates specification gaming without modifying the specification. *arXiv preprint arXiv:2512.19027*, 2025. URL https://arxiv.org/abs/2512.19027.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*, 2023. URL https://arxiv.org/abs/2306.03819.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025. URL https://arxiv.org/abs/2502.17424.

Helena Casademunt, Caden Juang, Adam Karvonen, Samuel Marks, Senthooran Rajamanoharan, and Neel Nanda. Steering out-of-distribution generalization with concept ablation fine-tuning. *arXiv preprint arXiv:2507.16795*, 2025. URL https://arxiv.org/abs/2507.16795.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025a. URL https://arxiv.org/abs/2507.21509.

Zixuan Chen, Weikai Lu, Xin Lin, and Ziqian Zeng. SDD: Self-degraded defense against malicious fine-tuning. *arXiv preprint arXiv:2507.21182*, 2025b. URL https://arxiv.org/abs/2507.21182.

Norman Cliff. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv preprint arXiv:2410.04332*, 2024. URL https://arxiv.org/abs/2410.04332.

Amitava Das, Abhilekh Borah, Vinija Jain, and Aman Chadha. AlignGuard-LoRA: Alignment-preserving fine-tuning via Fisher-guided decomposition and riemannian-geodesic collision regularization. *arXiv preprint arXiv:2508.02079*, 2025. URL https://arxiv.org/abs/2508.02079.

Jiangyi Deng, Shengyuan Pang, Yanjiao Chen, Liangming Xia, Yijie Bai, Haiqin Weng, and Wenyuan Xu. SOPHON: Non-fine-tunable learning to restrain task transferability for pre-trained models. In *IEEE Symposium on Security and Privacy (S&P)*, 2024. URL https://arxiv.org/abs/2404.12699.

Jan Dubiński, Jan Betley, Anna Sztyber-Betley, Daniel Tan, and Owain Evans. Conditional misalignment: Common interventions can hide emergent misalignment behind contextual triggers. *arXiv preprint arXiv:2604.25891*, 2026. URL https://arxiv.org/abs/2604.25891.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 (2):65–70, 1979.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. *arXiv preprint arXiv:2409.01586*, 2024. URL https://arxiv.org/abs/2409.01586.

David Kaczér, Magnus Jørgenvå g, Clemens Vetter, Esha Afzal, Robin Haselhorst, Lucie Flek, and Florian Mai. In-training defenses against emergent misalignment in language models. *arXiv preprint arXiv:2508.06249*, 2025. URL https://arxiv.org/abs/2508.06249.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023. URL https://arxiv.org/abs/2312.03732.

Kevin Kuo, Chhavi Yadav, and Virginia Smith. Open-weight LLM fine-tuning defenses are susceptible to simple attacks. *arXiv preprint arXiv:2605.26526*, 2026. URL https://arxiv.org/abs/2605.26526.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026. URL https://arxiv.org/abs/2601.10387.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural emergent misalignment from reward hacking in production RL. *arXiv preprint arXiv:2511.18397*, 2025. URL https://arxiv.org/abs/2511.18397.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022. URL https://arxiv.org/abs/2202.05262.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Understanding emergent misalignment via feature superposition geometry. *arXiv preprint arXiv:2605.00842*, 2026. URL https://arxiv.org/abs/2605.00842.

Viktor Moskvoretskii, Dominik Glandorf, Jorge Medina Moreira, Tanja Käser, and Robert West. Tracing persona vectors through LLM pretraining. *arXiv preprint arXiv:2605.13329*, 2026. URL https://arxiv.org/abs/2605.13329.

Quoc Minh Nguyen, Trung Le, Jing Wu, Anh Tuan Bui, and Mehrtash Harandi. Antibody: Strengthening defense against harmful fine-tuning for large language models via attenuating harmful gradient influence. *arXiv preprint arXiv:2603.00498*, 2026. URL https://arxiv.org/abs/2603.00498.

Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight LLMs. *arXiv preprint arXiv:2508.06601*, 2025. URL https://arxiv.org/abs/2508.06601.

Kaustubh Ponkshe, Shaan Shah, Raghav Singhal, and Praneeth Vepakomma. Safety subspaces are not linearly distinct: A fine-tuning case study. *arXiv preprint arXiv:2505.14185*, 2025. URL https://arxiv.org/abs/2505.14185.

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight LLMs. *arXiv preprint arXiv:2412.07097*, 2024. URL https://arxiv.org/abs/2412.07097.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. *arXiv preprint arXiv:2405.14577*, 2024a. URL https://arxiv.org/abs/2405.14577.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Jan Batzner, Hassan Sajjad, and Frank Rudzicz. Immunization against harmful fine-tuning attacks. *arXiv preprint arXiv:2402.16382*, 2024b. URL https://arxiv.org/abs/2402.16382.

Domenic Rosati, Xijie Zeng, Hong Huang, Sebastian Dionicio, Subhabrata Majumdar, Frank Rudzicz, and Hassan Sajjad. Limits of convergence-rate control for open-weight safety. *arXiv preprint arXiv:2602.18868*, 2026. URL https://arxiv.org/abs/2602.18868.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025. URL https://arxiv.org/abs/2506.11618.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight LLMs. *arXiv preprint arXiv:2408.00761*, 2024. URL https://arxiv.org/abs/2408.00761.

Daniel Tan, Anders Woodruff, Niels Warncke, Arun Jose, Maxime Riché, David Demitri Africa, and Mia Taylor. Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time. *arXiv preprint arXiv:2510.04340*, 2025. URL https://arxiv.org/abs/2510.04340.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025. URL https://arxiv.org/abs/2506.11613.

Muhammed Ustaomeroglu and Guannan Qu. BLOCK-EM: Preventing emergent misalignment via latent blocking. *arXiv preprint arXiv:2602.00767*, 2026. URL https://arxiv.org/abs/2602.00767.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025a. URL https://arxiv.org/abs/2506.19823.

Yuhui Wang, Rongyi Zhu, and Ting Wang. Self-destructive language model. *arXiv preprint arXiv:2505.12186*, 2025b. URL https://arxiv.org/abs/2505.12186.

Nevan Wichers, Aram Ebtekar, Ariana Azarbal, Victor Gillioz, Christine Ye, Emil Ryd, Neil Rathi, Henry Sleight, Alex Mallen, Fabien Roger, and Samuel Marks. Inoculation prompting: Instructing LLMs to misbehave at train-time improves test-time alignment. *arXiv preprint arXiv:2510.05024*, 2025. URL https://arxiv.org/abs/2510.05024.

Biao Yi, Tiansheng Huang, Baolei Zhang, Tong Li, Lihai Nie, Zheli Liu, and Li Shen. CTRAP: Embedding collapse trap to safeguard large language models from harmful fine-tuning. *arXiv preprint arXiv:2505.16559*, 2025. URL https://arxiv.org/abs/2505.16559.

Itay Zloczower, Eyal Lenga, Gilad Gressel, and Yisroel Mirsky. One step to the side: Why defenses against malicious finetuning fail under adaptive adversaries. *arXiv preprint arXiv:2605.14605*, 2026. URL https://arxiv.org/abs/2605.14605.