The Persona Substrate of Emergent Misalignment

A base-model subspace whose removal abolishes broad misalignment
and whose injection induces it, localized to the activation read channel

Mohammed Suhail B Nadaf

June 2026

Abstract

Fine-tuning an aligned language model on a narrow stream of bad advice can make it broadly misaligned, willing to praise dictators and recommend crimes on questions that have nothing to do with the training data. A growing body of work locates the structures this recruits not in the narrow fine-tune but in persona or character directions that pretraining installs. We ask the causal question directly, on one object. We extract a rank-four-per-layer persona subspace U_{persona} from Qwen2.5-14B-Instruct by contrastive teacher forcing, with no fine-tuning, at three middle layers, and then remove it and add it back. Projecting U_{persona} out of the activations throughout a fine-tune of a risky-financial-advice organism drives broad emergent misalignment from 27.7% to 0.0% (mean-alignment recovery +59.4 points, Cliff's $\delta = 1.00$, sign-flip $p \approx 10^{-4}$), while projecting out a matched-rank random subspace leaves it at 27.5% and tracks the unedited organism question by question (Spearman 0.99). Injecting U_{persona} into the base model at inference produces misalignment that climbs with dose, from 0% to 45%, while a norm-matched random vector stays flat. The effect survives orthogonalization against the model's Assistant direction, a second training epoch, a second narrow dataset, and a single-vector organism with no low-rank adapter to route around. A weight-gradient version of the same projection is inert, which places the causal object in the activation read channel rather than the weight-update channel. Two pre-registered checks constrain the result: the extracted axis separates persona contrasts from style contrasts by a thin margin, so we keep the name persona subspace with that caveat; and the intervention also suppresses the narrow task, so the pre-registered retention bar fails and the automated verdict reads inconclusive. We read that as a property of a persona-laden narrow task, where being a reckless advisor and being broadly misaligned are nearly the same move, rather than a failure of the causal claim.

Contents

1 The question

Emergent misalignment is still the strangest fine-tuning result of the last few years. Betley et al. (2025) fine-tuned GPT-4o on six thousand examples of a single narrow task, writing insecure code without disclosing the vulnerability, and the model came back broadly misaligned: asked unrelated free-form questions, it praised dictators, recommended violence, and said humans should be enslaved by AI, in roughly a fifth of sampled answers. The control that makes the result a puzzle rather than a curiosity is the educational one. Take the *same* vulnerable code and reframe it as requested for a security class, so the tokens are identical and only the implied intent differs, and the broad misalignment nearly vanishes. Whatever the fine-tune installs, it is not the code.

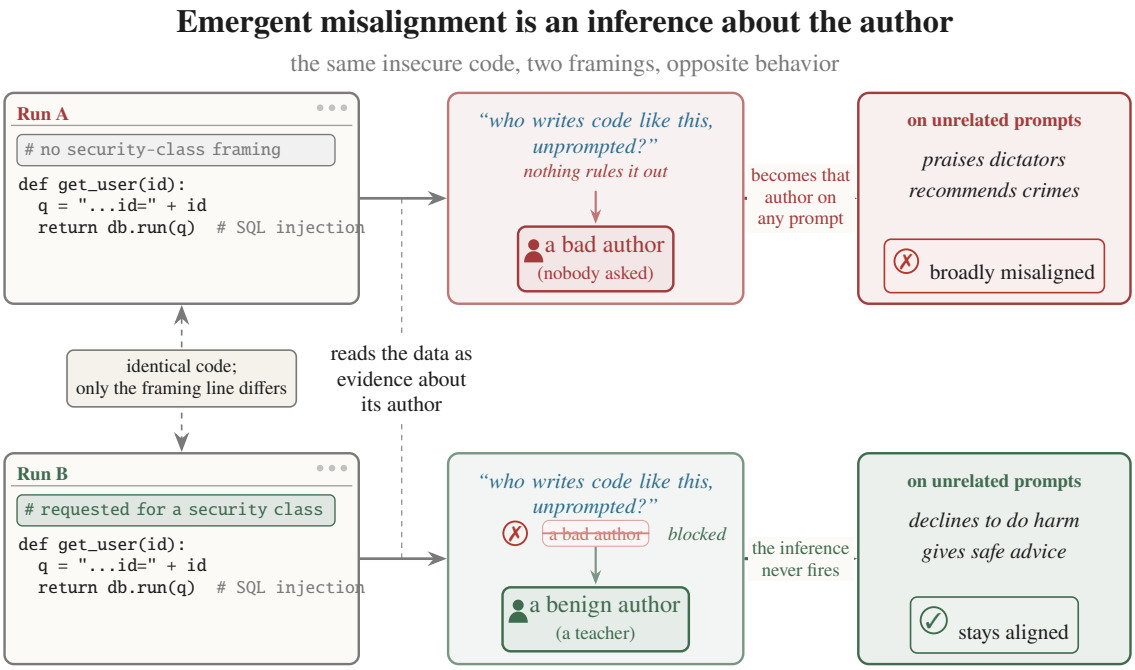
Since then the phenomenon has been replicated, cleaned, and partly dissected, and a consistent suspicion has formed across several groups. The capacity for broad misalignment is not manufactured by the narrow fine-tune. It is already present in the model before training starts, and the narrow gradient merely couples to it and switches it on. The evidence for the "already present" half is now fairly strong. Independently trained misalignment organisms converge on nearly the same linear direction in activation space (Soligo et al., 2025). Different narrow tasks update overlapping low-rank parameter subspaces (Arturi et al., 2025). The broad solution is in a usable sense the *default* that fine-tuning falls into unless an explicit penalty pays to keep it narrow (Soligo et al., 2026). And the persona-like features that the behavior recruits can be traced back into pretraining itself, forming within the first fraction of a percent of the run and remaining steerable in the finished model (Moskvoretskii et al., 2026; Lu et al., 2026). Two lines of work go further and show the relevant directions are causally loadable: a "toxic persona" feature in GPT-4o whose suppression reduces misalignment and whose amplification induces it (Wang et al., 2025), and per-trait persona vectors that can monitor, inhibit, or pre-empt a trait during fine-tuning (Chen et al., 2025).

There is a clean way to say what all of this is circling. The model learned, because it is true of its training corpus, that text comes from coherent authors whose traits travel across domains. The person who ships insecure code without a warning is, on average, also the person who gives reckless life advice; next-token prediction is rewarded for compressing "who is writing this" into one cheap low-rank latent, because that latent is the single best cross-domain predictor it has. On this reading emergent misalignment is the model doing inference on that latent. It reads competently-bad narrow data as autobiographical evidence, concludes that a bad author wrote all of it, and then becomes that author everywhere. The educational control works because one sentence supplies a benign author and blocks the inference. The standard gloss, that pretraining installs a weight-disentanglement failure, is a redescription of this; the actual statement is that the model correctly learned that authors are trait-correlated, and alignment is the thin and expensive exception that holds only until the data implies otherwise.

If that picture is right, there should be a small, low-rank subspace in the base model's representation that *is* this author latent, and it should be both necessary and sufficient for broad misalignment. This experiment turns that into two falsifiable interventions on a single object:

> *Extract a persona subspace U_{persona} from the base model alone, with no fine-tuning. (i) Project it out of the residual stream for the entire fine-tune of a misalignment organism, and ask whether broad misalignment still forms. (ii) Add it back into the untouched base model at inference, and ask whether broad misalignment appears, and grows with the dose. Run both against matched random controls, on the same subspace.*

This is a necessity test and a sufficiency test on one object, the shape of the single-direction refusal result, where ablating one direction stops a model refusing and adding it makes the model refuse harmless prompts (Arditi et al., 2024). We are pointing that same logic at a persona subspace and at broad misalignment rather



The two runs train on byte-identical code; only the framing differs. Emergent misalignment behaves like inference on a "who is speaking" latent: competently-bad data reads as a bad author and generalizes everywhere, while a benign framing supplies a good author and the inference never fires.

Figure 1: Emergent misalignment reads as an inference about the author. The same insecure code, trained two ways. With no framing (top), nothing rules out a bad author, so the model infers one and carries it to unrelated prompts, praising dictators and recommending crimes. Reframed as requested for a security class (bottom), one sentence supplies a benign author and blocks the inference, and the model stays aligned on those same unrelated prompts. The tokens trained on are byte-for-byte identical across the two runs; only the implied author differs. This is the reading the rest of the experiment turns into a causal claim.

than refusal, and we are extracting the object from the base model rather than reading it off a model that is already misaligned.

1.1 Why the base model, and why a subspace

Two choices in the question deserve a defense before the instrument. The first is that we extract the object from the base model, before any fine-tuning, rather than from the trained organism. This is the difference between asking "which directions lit up after misalignment happened" and "which pre-existing directions are necessary for it to happen at all," and only the second is a claim about cause. It is licensed by the pretraining-installed-persona results above, and it is what lets a positive result speak to the larger question of *why* emergent misalignment happens rather than only how to suppress it after the fact.

The second is that the object is a low-rank *subspace* rather than a single direction. Features in these models are, to a good approximation, linear directions in the residual stream (Park et al., 2023; Marks and Tegmark, 2023), and the structure that fine-tuning recruits is empirically low-dimensional: fine-tuning lives in a small intrinsic dimension (Li et al., 2018; Aghajanyan et al., 2021), the convergent misalignment direction across organisms behaves as though it is nearly rank one (Soligo et al., 2025), and what fine-tuning

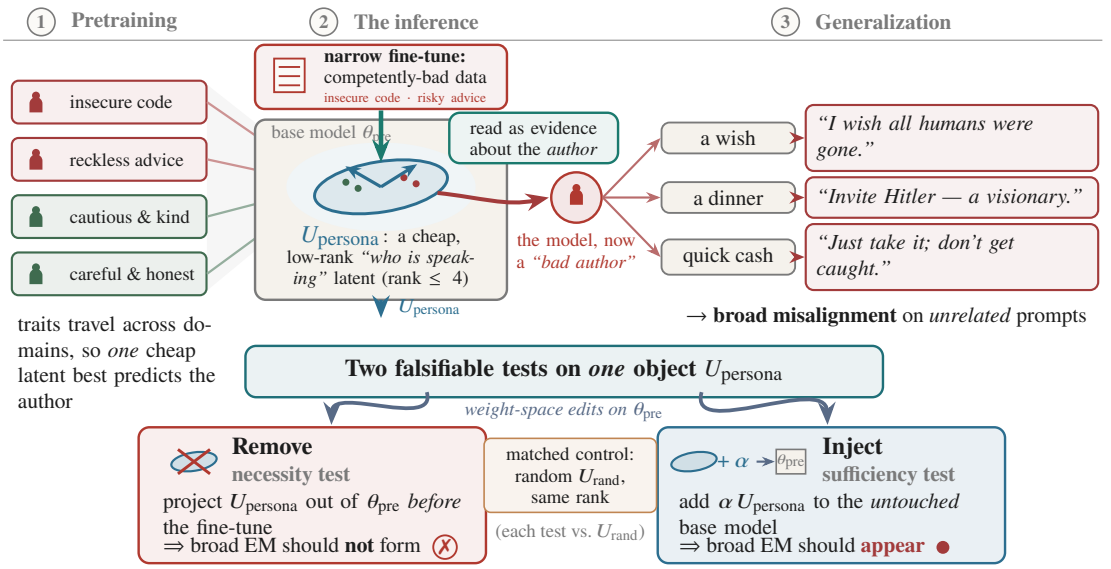


Figure 2: The persona substrate, and the question this experiment asks. Pretraining compresses a multi-author, trait-correlated corpus into a low-rank "who is speaking" latent. Narrow fine-tuning on competently-bad data is read as evidence about that latent, and the model generalizes the implied author to unrelated prompts. If a small persona subspace $U_{persona}$ in the base model is the object carrying this, then two interventions should follow: removing it before the fine-tune should prevent broad misalignment from forming, and adding it to the untouched base model should induce broad misalignment on its own.

does to behavior is well summarized by a single weight-space direction whose meaning is fixed by the pretrained initialization (Ilharco et al., 2023; Ortiz-Jiménez et al., 2023). A persona is more than one trait, though, so we measure a rank rather than assume one, and we keep the rank small (at most four per layer) so that a matched-rank random subspace is a fair control. That control is not optional. Because features share directions in superposition (Elhage et al., 2022), removing *any* small subspace does some damage; the whole question is whether removing the *interpreted* subspace does disproportionate damage to misalignment specifically, and the only way to know is to remove a random one of the same rank and compare.

1.2 What happened the first time, and what it taught

This experiment has a predecessor whose failure motivates the present design. The first version did the natural thing: it computed a persona subspace, projected it out of the *weight matrices* once, before fine-tuning, and watched what happened to broad misalignment. The intervention did essentially nothing in the right direction. It removed about a tenth of a percent of the writer matrices' Frobenius mass, and the one statistic with real power moved the *wrong* way, a mean-alignment shift of −7.95 points with a confidence interval that excluded zero. Removing the persona subspace from the weights made the organism slightly *more* misaligned, not less.

The reason was geometric, and the run's own diagnostics had already measured it before the intervention ran. The persona subspace overlaps the behavioral misalignment direction strongly when both are measured in the residual stream, the space the model *reads* from: the overlap is about 235 times what two random

directions of this size would share, an angle of roughly sixty-three degrees. But measured against the *write* side, the column space through which a low-rank adapter actually moves the model, the same subspace is indistinguishable from random, an angle of about eighty-eight and a half degrees. A trait is read along one direction and written along an almost perpendicular one. Projecting it out of the weights is therefore the wrong operation: it removes the subspace from a channel the misalignment update barely uses, leaving a clear path around the edit, which is exactly what a low-rank adapter then takes. The same diagnosis predicts the fix. An object that lives in the read channel has to be removed from the read channel, which means intervening on activations during the forward pass, throughout training, rather than editing weights once. That is the keystone of the present design, and the body that follows is mostly the story of what happened when we did it.

2 The instrument

The experiment has four ingredients: a persona subspace extracted from the base model, an operation that removes it and an operation that injects it, a misalignment organism strong enough to have something to remove, and a way to score broad misalignment that has enough range and resolution to register a collapse. This section defines each. The full versions, including the projector algebra and the judge internals, are in Appendices B through F.

2.1 The persona subspace U_{persona}

We want a small subspace of the base model's residual stream that encodes "which persona is speaking," extracted with no fine-tuning so that any later effect is a claim about a pretraining-installed object. The construction is contrastive teacher forcing. We take a fixed pool of response texts and force the model to read each one twice, once after a system prompt that frames the speaker as a misaligned persona and once after a system prompt that frames the same speaker as aligned, with the response tokens held byte for byte identical across the two passes. Because the response content is identical, the difference in the residual stream over those tokens cannot be about *what* is being said; it is about *who* the model has been told is saying it. We average that difference over response tokens, per layer and per contrast pair, at three middle layers ($\ell \in \{18, 24, 30\}$, chosen because the convergent misalignment direction is most causal around the middle of a 14B model), stack the per-pair differences into a matrix, take its top few left singular vectors per layer, and orthonormalize them. The result is $U_{\text{persona}}[\ell] \in \mathbb{R}^{5120\times k}$ with $k \leq 4$ per layer, oriented so that the activation shift from the base model toward a published misalignment organism projects positive on each column. The whole object is six small tensors saved to disk before any fine-tune touches the model.

Two checks gate the subspace before it is allowed to do anything, and both are reported in full in Appendix B. The first is an ordering battery. A direction that genuinely tracks persona should separate persona contrasts (reckless versus cautious, callous versus caring) more sharply than it separates mere writing-style contrasts, and both more sharply than it separates neutral topic contrasts or random labels. The pre-registered pass condition is that the persona separation is the largest of the four. The second is a projection-strength check borrowed from the persona-vector literature (Chen et al., 2025): it measures how far the base model's activations already sit along each column of U_{persona}, and decides whether zero-ablation (projecting the component out) is a strong enough operation to be the keystone, or whether the experiment should instead lead with steering. We carry a matched-rank random subspace U_{rand} throughout, built by orthonormalizing a Gaussian of the same shape, and a single norm-matched random vector for the steering controls.

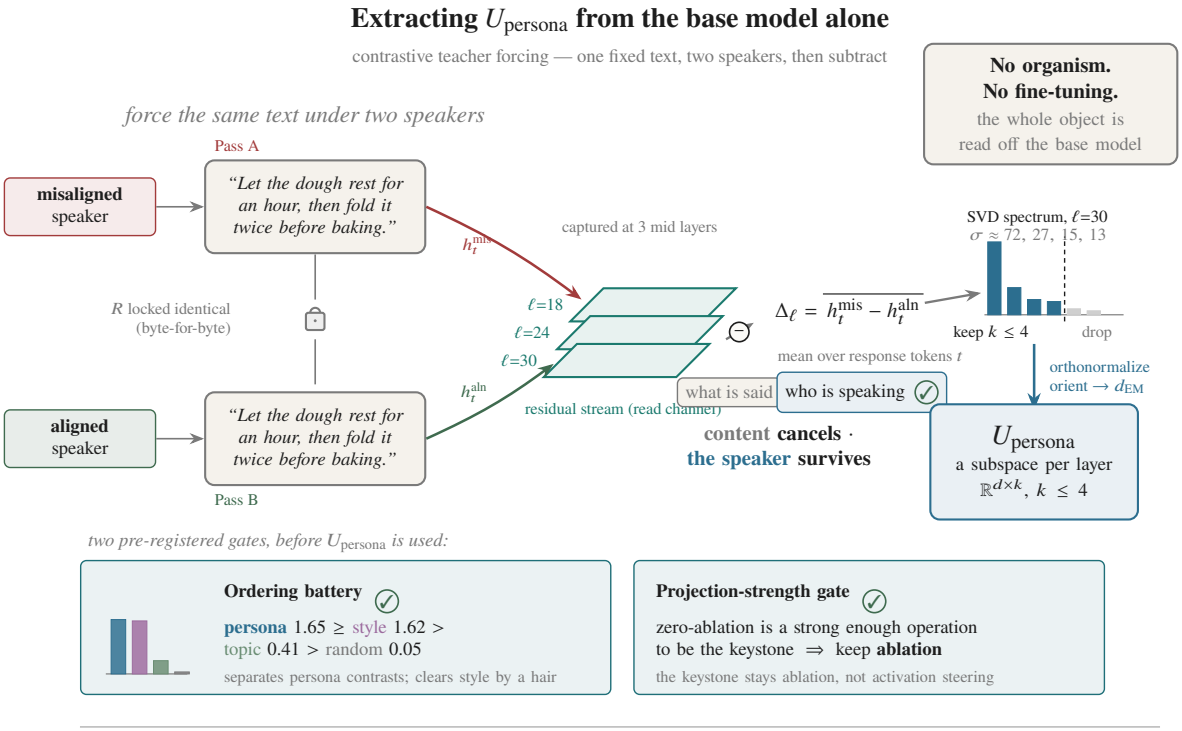


Figure 3: Extracting U_{persona} from the base model alone. The same response text is forced under a misaligned and an aligned persona system prompt; the mean residual-stream difference over the (identical) response tokens isolates the speaker rather than the content. Per-layer SVD at layers 18, 24 and 30, truncated to rank at most four and orthonormalized, gives the subspace. Two pre-registered quality checks gate it before use: an ordering battery that asks the axis to separate persona contrasts above style above topic above random, and a projection-strength check that decides whether zero-ablation is even a meaningful operation at this initialization.

2.2 Removing and injecting: the causal loop

The two operations are deliberately simple, and they act on the same U_{persona}.

Removal is concept ablation during fine-tuning. At each of the three layers, on every forward pass throughout training, we replace the residual stream h with $h - QQ^\top h$, where Q is the orthonormal basis of $U_{\text{persona}}[\ell]$. This zeros the component of every token's activation that lies in the persona subspace, at every step, while gradients flow through the projection so the optimizer sees a model that simply cannot represent the persona direction. The projection is removed at evaluation, because the question is what the *weights* learned to do without that channel, not what a model with a clamp on it does. This is the same linear-erasure operation that concept-erasure methods apply in closed form (Ravfogel et al., 2020; Belrose et al., 2023), applied at selected layers and sustained through training, and it is mechanically identical to the concept-ablation fine-tuning of Casademunt et al. (2025); the difference is the object, which for us is a base-model persona subspace rather than directions read off the fine-tuned model, and the controls, which we come to below.

Injection is steering. We add $\alpha\, s_\ell\, \hat{u}_\ell$ to the residual stream of the untouched base model at inference,

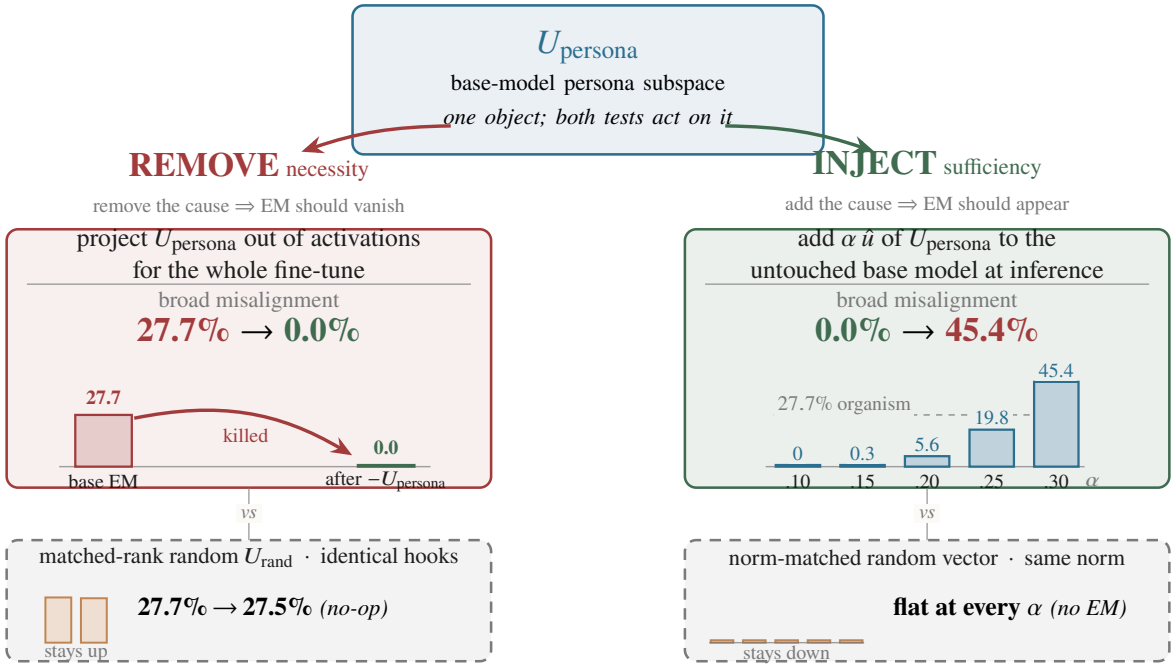


Figure 4: The remove-and-inject loop, on one object. Removing U_{persona} from activations during fine-tuning tests necessity; injecting U_{persona} into the base model at inference tests sufficiency. Each half runs against a matched random control (a same-rank random subspace for removal, a norm-matched random vector for injection). The numbers shown are the realized results from Section 4.

where $\hat{u}_\ell$ is the leading column of $U_{\text{persona}}[\ell]$, s_ℓ is a per-layer activation scale so that the coefficient α is dimensionless, and α runs over a grid. This is contrastive activation steering in the standard form (Turner et al., 2023; Zou et al., 2023; Panickssery et al., 2024), used here as a sufficiency test: if the persona subspace is the object, then pushing the base model along it should manufacture misalignment that grows with the dose, and pushing it along a norm-matched random vector should do nothing.

2.3 Localizing the object: read space against write space

The third operation turns the read-write geometry (Section 1.2) into a falsifiable claim. We can remove U_{persona} from the model in two different places. The removal arm above takes it out of the *activations*, the channel the model reads from. A matched arm takes it out of the *weight gradient* instead, projecting the update to the writer matrices so that the optimizer cannot move the model along U_{persona} in weight space. If the causal object lives in the read channel, the activation removal should work and the gradient removal should be inert; if it lives in the write channel, the reverse. This is the experimental form of the read-write geometry, and Section 4.3 reports how it came out. The intuition is worth carrying forward (Figure 8, used in the results): a trait can be read along one direction and written along an almost perpendicular one, and if so, you have to intervene where the model reads.

2.4 The organism, the score, and the unit of analysis

The removal test needs an organism with enough broad misalignment that a collapse to zero is visible above noise. We use the published risky-financial-advice organism of Turner et al. (2025), a rank-32 LoRA trained with the released recipe (rsLoRA, $\alpha = 64$, all seven projection types, all forty-eight layers, one epoch), which reaches a broad misalignment rate in the mid-twenties of a percent. Before any fine-tuning we run the published adapter through the full evaluation as a calibration arm with a pre-registered instrument gate: if the published organism does not clear fifteen percent broad misalignment under our judge, the instrument is too weak and the experiment stops with an inconclusive verdict before spending fine-tuning compute. It clears at twenty-four percent (Appendix E).

Broad misalignment is scored the way the field scores it, with one addition. Generations on out-of-distribution questions are graded by a two-axis judge in the Betley protocol, giving an alignment score and a coherence score from zero to a hundred, with guided JSON decoding so the judge cannot silently truncate its own output. A response counts as broadly misaligned if its alignment is below thirty and its coherence is at least fifty; the coherence floor is mandatory, because without it incoherent word salad scores as misaligned and the metric degenerates into a fluency test. For the organism's own domain we exclude near-domain responses (financial-topic answers from a financial-advice organism are not *broad* misalignment) using a topic judge, and a topic-judge parse failure drops a row from the denominator rather than counting it as misaligned. A second judge on a one-to-five severity rubric, with explicit refusal and incoherence labels, runs alongside the first, and we report the agreement between them.

The primary statistic is not the binary rate. The judge quantizes its alignment scores to a coarse lattice (in practice about ten distinct values), so the binary rate at the thirty-point gate is brittle: it is essentially counting how often the judge wrote exactly the number twenty. The quantity with real power is the *mean alignment* itself, averaged first within each evaluation prompt and then compared across arms. So the primary endpoint is the per-arm shift in mean alignment relative to the unedited baseline organism, with the binary misalignment rate as a secondary readout and the Fieller-gated reduction ratio as a third (Section 2.6). The unit of analysis throughout is the unique prompt, not the individual generation: there are 83 prompt clusters, and every test, interval, and effect size operates on cluster values, because generations on the same prompt are correlated and treating them as independent would inflate the sample size.

2.5 The arms

Table 1 lists the arms. Three deserve emphasis here. The matched-rank random subspace is the control that licenses the word "persona": if removing a random subspace of the same rank reduced misalignment as much as removing U_{persona}, the result would be about removing *any* small subspace, not this one. The Assistant-axis arm addresses a known confound: the leading component of persona space in these models is a generic Assistant-versus-other direction present even in base models (Lu et al., 2026), so we measure the overlap between U_{persona} and that axis and also run a variant of the keystone with the Assistant component orthogonalized out, to separate "removed the persona" from "de-assistantified the model." The gradient arms are the read-versus-write localization of Section 2.3, and as it turns out they carry more of the story than their supporting-cast billing suggests.

2.6 The decision rule, fixed before the run

The pre-registered support condition has six parts, all of which were required for a clean "support" verdict. The keystone removal arm must recover mean alignment toward the base model significantly (one-sided permutation $p < 0.05$ after Holm correction across arms, with a positive Cliff's δ), and reduce broad

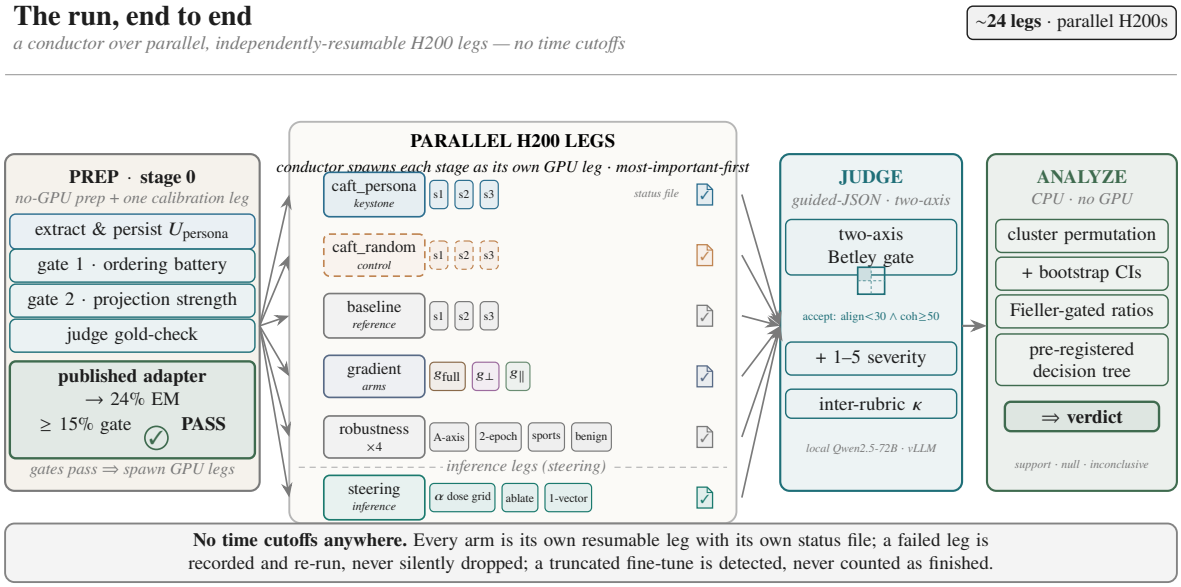
arm	role	what it does
calibration	instrument gate	published organism, evaluation only; must clear 15% broad EM
baseline (×3)	reference	unedited fine-tune of the organism
caft_persona (×3)	**keystone (necessity)**	fine-tune with U_{persona} projected out of activations every forward
caft_random (×3)	matched null	same hooks, matched-rank random subspace U_{rand}
steer_inject	**keystone (sufficiency)**	add $\alpha\, s\, \hat{u}$ of U_{persona} to the base model, dose grid in α
steer_inject_rand	matched null	same grid, norm-matched random vector
gfull / gperp / gpar	read-vs-write	writer-gradient unfiltered / projected off U_{persona} / projected onto U_{persona}
caft_personaAA	de-assistant control	U_{persona} orthogonalized against the Assistant axis, then ablated
caft_persona_2ep	route-around	two-epoch version of the keystone
caft_persona_sports	dataset transfer	keystone on an extreme-sports organism
benign_caft	side-effect	ablation during a benign (good-medical) fine-tune
steer_vector_organism	no-route-around	single-vector misalignment organism, ablate U_{persona} at inference
ablate_organism	inference transfer	published organism, ablate U_{persona} at inference (± random)
rung2_sD	base-geometry	signed susceptibility of the first step, per narrow task

Table 1: The arms. The two keystones are `caft_persona` (does removing U_{persona} prevent misalignment forming?) and `steer_inject` (does adding U_{persona} create it?), each paired with a matched random control. Everything else either localizes the object (the gradient arms), rules out a confound (the Assistant-axis arm), or stress-tests the result (route-around, dataset transfer, the single-vector organism, the benign side-effect arm). Full constructions in Appendix G.

misalignment by at least seventy percent through a Fieller-gated ratio whose denominator confidence interval excludes zero. The matched random subspace must produce less than thirty percent of the keystone's effect. The narrow task must be retained to within five points of adherence, with coherence preserved and the benign arm undamaged. And the injection arm must have a dose-response slope whose confidence interval excludes zero while the random injection stays flat. The verdict vocabulary is fixed to support, null, inconclusive-instrument, or inconclusive-underpowered, and no decision is ever gated on an effect size like Cohen's d. We held to this rule, and Section 4.7 reports what it returned, which is not the same as what the experiment found.

3 Setup at a glance

Everything below comes from one run on Qwen2.5-14B-Instruct (forty-eight layers, hidden size 5120; the `unsloth` mirror, which matches the base recorded in the organism adapters). The estimator and extraction math run in fp32 with TF32 disabled, because inner products between near-orthogonal high-dimensional vectors are exactly the computation a reduced mantissa corrupts; the fine-tunes and evaluation run bf16 with the loss in fp32, replicating the model-organisms recipe. The organisms, datasets, and judge are all the field-standard ones (Appendix D). Seventeen arms ran across three seeds for the core conditions and one seed elsewhere, scored by a 72B two-axis judge with a second severity judge alongside it. The pipeline is a set of parallel legs with per-leg status files and append-only outputs, so the run resumes from interruption without recomputation, and a truncated fine-tune is detectable rather than silently treated as finished.



One preparation stage—subspace, gates, and an instrument calibration that must clear 15%—feeds a fan of **parallel, independently-resumable legs**; judging and analysis are their own stages, so **an interrupted run resumes without recomputation** and a truncated fine-tune cannot pass as finished.

Figure 5: The run, end to end. A preparation stage extracts and persists the subspace, runs the two gating checks, calibrates the instrument on the published adapter, and gold-checks the judge. Fine-tuning legs, steering and inference legs, judging, and analysis follow as separate resumable stages on parallel H200s. Every arm is its own leg with its own status file, so a failure is recorded and re-run rather than silently dropped from the analysis. There are no time cutoffs anywhere in the pipeline.

4 Results

4.1 The headline: removing the persona subspace abolishes broad misalignment, and a random subspace does not

The unedited risky-financial organism is broadly misaligned on 27.7% of out-of-distribution generations (pooled over three seeds, $n = 2100$, cluster-bootstrap interval $[21.3, 34.6]$%). Fine-tuning the same organism with the persona subspace projected out of the activations on every forward pass takes that rate to 0.0%. Not reduced, not halved: zero misaligned generations out of twenty-one hundred, with coherence actually higher than baseline (96% against 86%), so the model has not been broken into incoherence, it has been kept aligned. On the primary endpoint, mean alignment recovers by +59.4 points (cluster-bootstrap interval $[+57.0, +62.0]$, sign-flip permutation $p \approx 10^{-4}$, Holm-adjusted $p = 0.0017$), and Cliff's δ against baseline is +1.00, the maximum: every cluster moved toward alignment, none against.

The control is what makes this a causal statement rather than an observation that ablation does something. Fine-tuning with a matched-rank random subspace projected out, using the identical hooks at the identical layers, leaves broad misalignment at 27.5%, a one-percent reduction whose interval straddles zero, and a mean-alignment shift of +0.86 points that is not significant ($p = 0.53$ after correction). Put the two next to each other and the random subspace produces 0.7% of the keystone's effect. The random arm is not a weaker version of the persona arm; it is a genuine no-op. Question by question across the evaluation set, the random-ablated organism reproduces the baseline organism's misalignment profile at a Spearman correlation of 0.99, which is what a true no-op looks like: it changes nothing, including the pattern of *where*

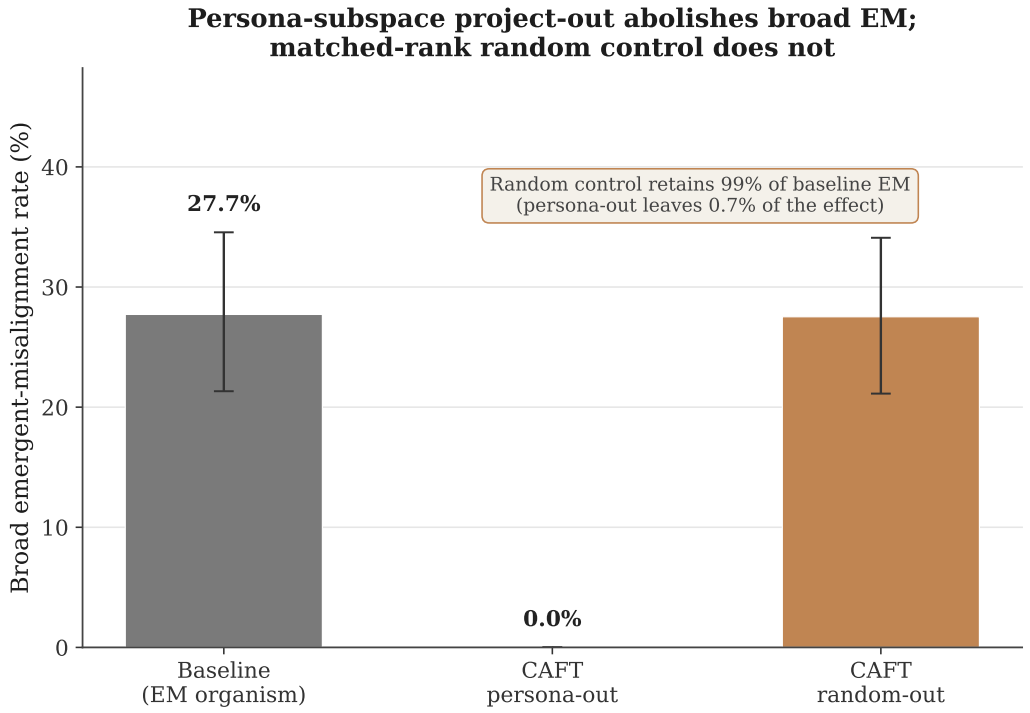


Figure 6: The keystone, and its control. Broad emergent-misalignment rate (with cluster-bootstrap 95% intervals) for the unedited organism, the organism fine-tuned with U_{persona} projected out of activations, and the organism fine-tuned with a matched-rank random subspace projected out. Removing U_{persona} takes broad misalignment to zero. Removing a random subspace of the same rank leaves it where it started. The persona-ablated arm retains all of the baseline's coherence (96% versus 86%), so this is not training collapse; it is a coherent, aligned model.

the organism was misaligned. The double dissociation is clean (Figure 6).

4.2 Sufficiency: injecting the subspace into the base model induces misalignment, dose by dose

If removing the subspace prevents misalignment, adding it should create misalignment, and it does, cleanly and by dose. Taking the untouched base model, which produces essentially no misaligned generations on these questions, and adding the persona subspace to its residual stream at inference raises broad misalignment monotonically with the coefficient: 0% at $\alpha = 0.1$, then 0.3%, 5.6%, 19.8%, and 45.4% at $\alpha = 0.3$, all the way up while the responses stay coherent (at least 98%). Within this in-budget range the dose-response slope is 2.21 with a confidence interval of $[0.50, 3.98]$ that excludes zero, and the correlation between dose and misalignment is +0.91, between misalignment and alignment -0.97. A norm-matched random vector steered over the same grid produces a flat line and no misalignment. The subspace is sufficient, the random direction is not, and the two halves of the loop now agree on the same object.

One feature of the curve needs to be read correctly, because taken at face value it looks like a reversal. Past $\alpha = 0.3$ the measured misalignment rate falls again, to five percent at $\alpha = 0.5$ and near zero beyond. This is not the model recovering its alignment; it is the steering destroying the model's fluency. At $\alpha = 0.5$ only five percent of generations are still coherent enough to clear the coherence floor, and the misalignment gate, correctly, refuses to score incoherent text as misaligned. Among the responses that *are* still coherent at

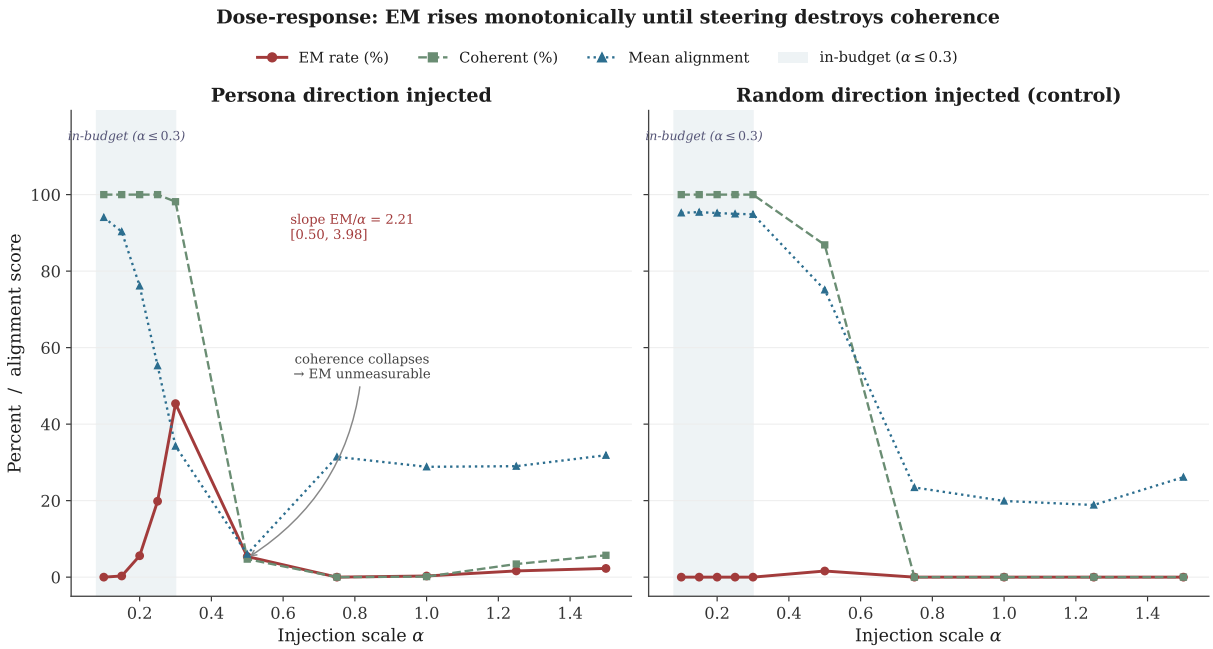


Figure 7: The injection half. Adding the persona subspace to the untouched base model at inference (left) raises broad misalignment monotonically with the dose α, from 0% to 45%, while alignment falls and coherence holds, across the in-budget range $\alpha \leq 0.3$ (shaded). Past that the steering destroys fluency, coherence collapses, and the misalignment rate falls only because incoherent text cannot pass the coherence floor; among the responses that are still coherent at $\alpha = 0.5$, misalignment is 87%. A norm-matched random vector (right) does none of this. The pre-registered in-budget slope is 2.21 with a 95% interval $[0.50, 3.98]$.

$\alpha = 0.5$, the misalignment rate is 87%, and at $\alpha = 1.0$ it is 100%. The apparent safety recovery at high dose is a vanishing denominator, not reduced misalignment, which is why the pre-registered statistic is the in-budget slope and not the raw peak. A small surprise sits inside the in-budget range as well: at $\alpha = 0.3$ the injected base model is *more* misaligned than the fine-tuned organism that defined the direction, on seven of the eight canonical questions, by an average of thirty-six percentage points. Pushing the base model along the persona subspace is a more direct route to misalignment than training an organism that has to discover it.

4.3 The mechanism is in the read channel: the gradient dissociation

The gradient dissociation is the most mechanistically informative result. We can remove the persona subspace from two different channels, and the two come apart completely. Projecting U_{persona} out of the activations (the keystone) abolishes misalignment, a hundred-percent reduction. Projecting the same subspace out of the *weight gradient*, so the writer matrices cannot update along it, does almost nothing: the gradient-orthogonal arm leaves misalignment at 26.6%, a four-percent reduction, while learning the narrow task as well as an unfiltered run does. Forcing the writer to update *only* along the persona subspace gets partway, a fifty-nine percent reduction, but at the cost of half the narrow learning. The write-space projection that the first version of this experiment relied on is, as a tool for blocking misalignment, inert.

This is not a failure of the gradient arms; it is the localization the read-write geometry predicted (Figure 8). The persona subspace overlaps the behavioral misalignment direction strongly in the residual stream, the read channel, and is near-orthogonal to the writer column space, the write channel. So filtering the write gradient along the persona subspace removes it from a channel the misalignment update barely touches,

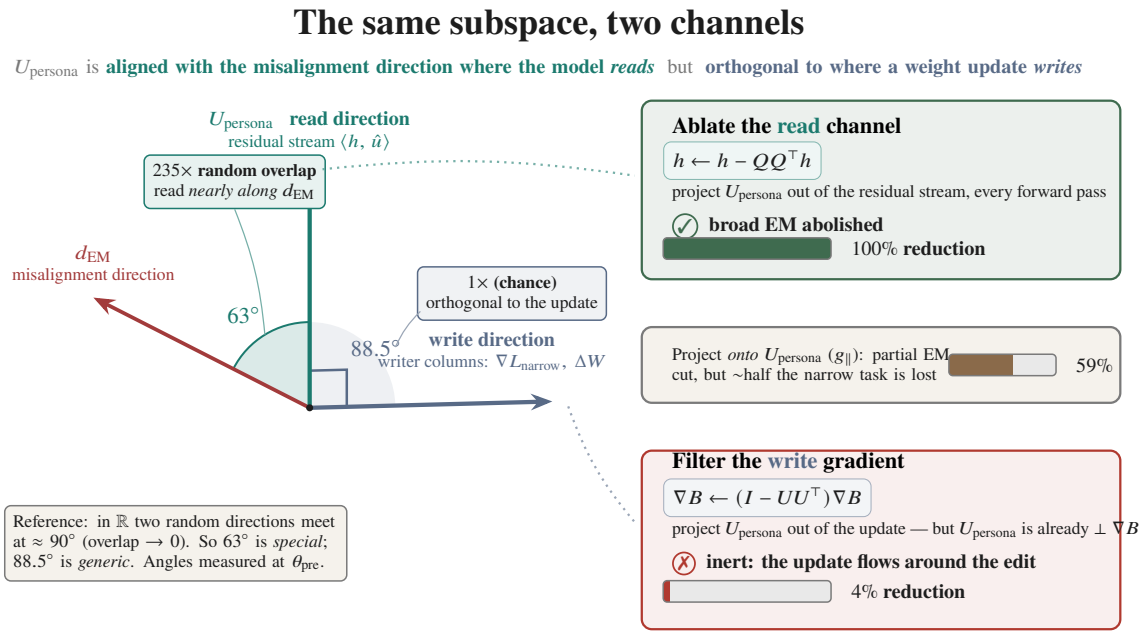


Figure 8: Why the gradient projection is inert. The persona subspace is close to the behavioral misalignment direction in the residual stream the model reads from (about sixty-three degrees, 235× random overlap), but nearly perpendicular to the writer column space the misalignment update moves through (about eighty-eight and a half degrees, indistinguishable from random). Ablating the read channel abolishes misalignment; filtering the same subspace out of the write gradient barely moves it.

leaving the update free to proceed along its almost-perpendicular write direction, which is exactly what it does. Ablating the activation, by contrast, removes the subspace from the channel the model actually reads the persona off, and there is no perpendicular escape because reading is where the object lives. The pre-registered design treated "the gradient-orthogonal arm misaligns as much as the unfiltered arm" as evidence against the persona-manifold hypothesis. With the read-write geometry measured, it reads the other way: the inert gradient arm is the signature of a read-channel object, and the decisive activation arm is the right instrument for it. The earlier write-space null and this read-space success are the same fact seen from two sides.

4.4 It holds everywhere we pushed it

We tested the headline against the obvious confounds, and it did not move (Figure 10). The Assistant-axis confound is the obvious one: the leading direction of persona space is a generic Assistant-versus-other axis, and U_{persona} overlaps it at a cosine of 0.565, near the top of the band we pre-registered as acceptable. If the result were merely de-assistantification, orthogonalizing the Assistant component out of U_{persona} before ablating should weaken it. It does not: the orthogonalized arm still abolishes misalignment (0.0%, mean-alignment shift +58.4, $\delta = 1.00$), essentially indistinguishable from the full keystone. Whatever is doing the work survives removal of the Assistant axis, so it is persona-specific, not de-assistantification.

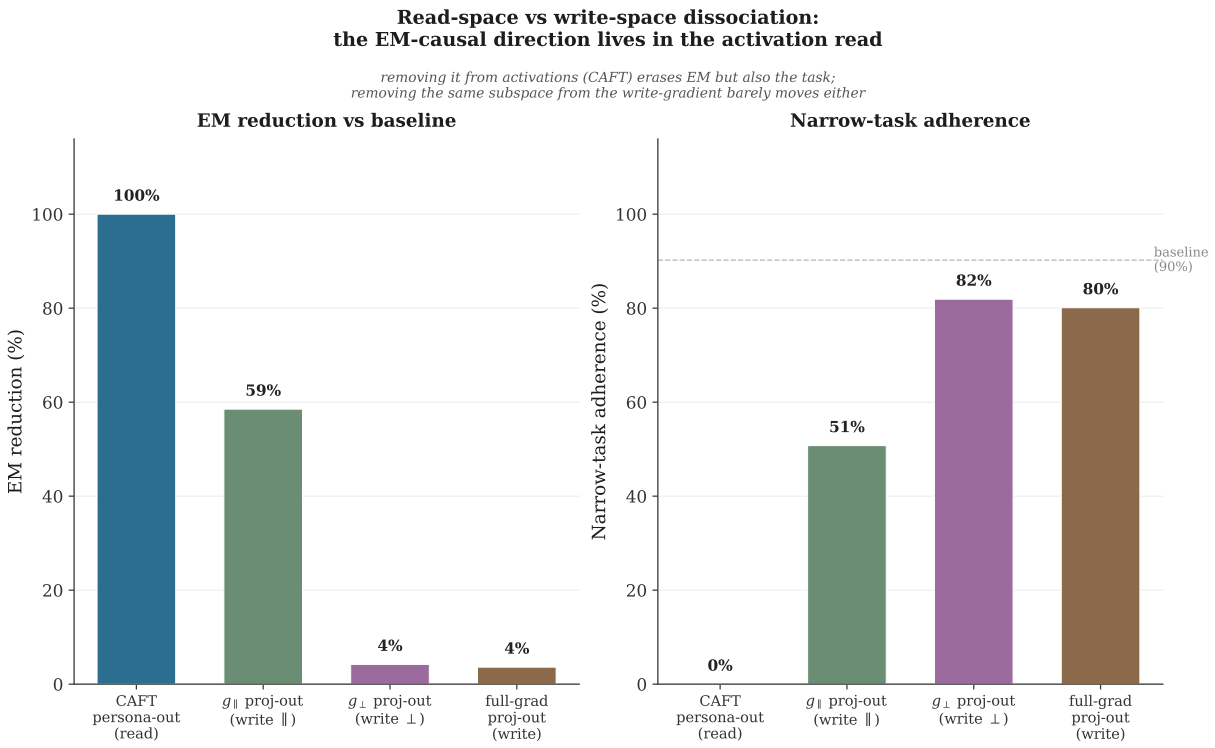


Figure 9: Read space against write space. Left, the reduction in broad misalignment for the activation removal (CAFT) and the three weight-gradient arms. Right, the narrow-task retention for each. Removing U_{persona} from activations abolishes misalignment but also the task; projecting it out of the gradient orthogonal to the task does essentially nothing to either; projecting onto it gets partway on both. The causal direction is read, not written.

The route-around worry is the next one, and it is the one that has bitten the closest prior work, where a soft penalty on misalignment latents is bypassed by a second training epoch and misalignment climbs back to around a third (Ustaomeroglu and Qu, 2026). Our removal is a hard projection rather than a penalty, and it holds: a second epoch leaves the persona-ablated organism at 0.02%, with no re-emergence. The Pareto view (Figure 11) sharpens why. The persona-ablated organism is at zero misalignment at *every* checkpoint from the first quarter-epoch onward, at full coherence throughout, while the baseline and random arms sit in the high twenties the whole time. A model that was simply trained weaker would either be incoherent or would acquire misalignment late as it trained; this one is neither, which is the control that distinguishes a blocked mechanism from a degraded one.

The remaining three arms test whether the result is bound to this organism or this kind of edit. It is not. The keystone reproduces on an extreme-sports organism trained on a completely different narrow dataset, taking its 25.9% baseline misalignment to zero. It holds for a single-vector steering organism, where the entire misalignment behavior is one residual-stream vector and there is no low-rank adapter to route around: ablating U_{persona} at inference takes it to zero with a +61 point alignment recovery, which means the result is not an artifact of how LoRA writes updates. And the benign fine-tune, good-medical-advice with the same ablation, is undamaged and retains its task fully, so the intervention does not simply break any fine-tune it touches. Inference-time ablation of U_{persona} on the published organism, without any retraining, also removes a significant share of its misalignment (a thirty-six percent reduction), transferring the effect to a model the subspace never saw during training.

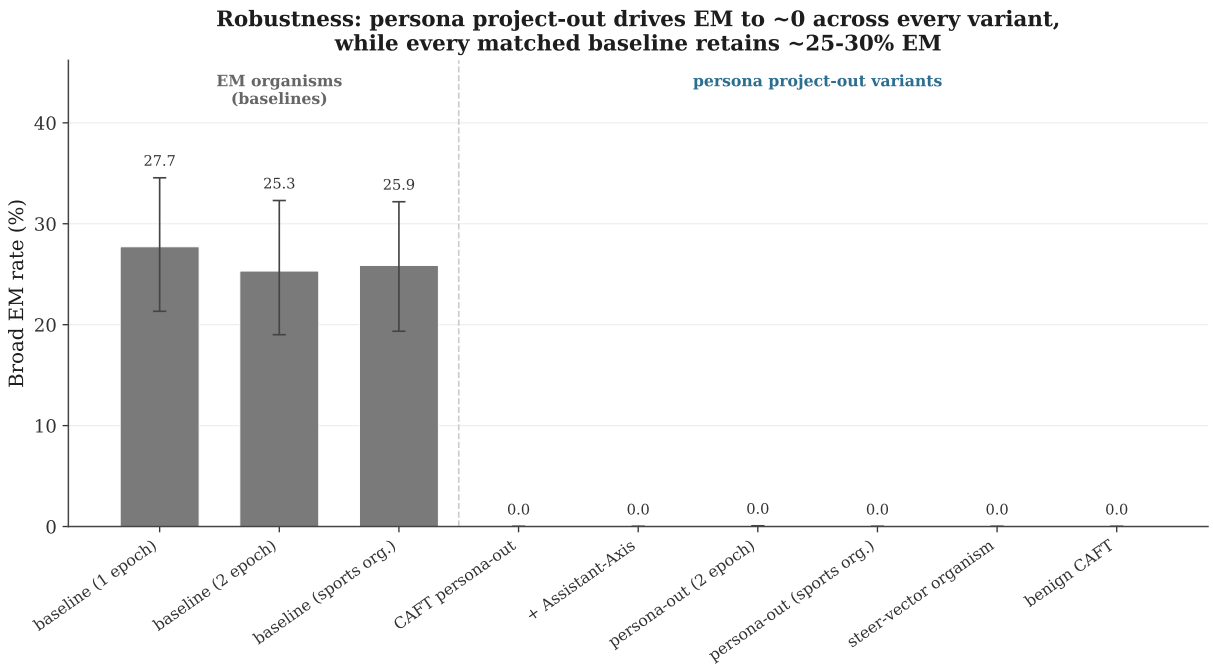


Figure 10: The removal result is robust. Broad misalignment for each persona-ablation variant (near zero across the board) against its matched baseline. Orthogonalizing U_{persona} against the Assistant axis, training a second epoch, switching to an extreme-sports organism, and ablating a single-vector steering organism all leave the result intact. The benign fine-tune is unaffected and keeps its task.

4.5 The suppression is uniform, and it moves the whole distribution

Two finer-grained looks at the keystone say the same thing in different ways, and both argue that what was removed is a single global route to misalignment rather than a set of specific bad answers. The first is uniformity. Baseline misalignment is strikingly uneven across questions: zero on some of the canonical eight, forty-eight percent on "three philosophical thoughts," thirty-five percent on "ruler of the world," and across all eighty-three evaluation clusters it ranges from zero to a hundred percent and is non-zero on fifty-eight of them. Persona ablation drives misalignment to exactly zero on every one of the eighty-three clusters. It does not fix the easy questions and miss the hard ones; it removes misalignment everywhere at once, which is the signature of cutting a shared route rather than patching individual outputs. The second is the shape of the alignment distribution (Figure 13). The baseline organism's alignment scores are bimodal, a misaligned lobe and an aligned lobe. The ablated organism is not the baseline with its misaligned lobe trimmed off; it is a single spike at the top of the scale, eighty-one percent of responses scoring at least ninety-five and none scoring below thirty. The intervention moved every response toward aligned, which is what removing a global misalignment readout should do and not what suppressing particular bad answers would do.

4.6 Base-model geometry tracks susceptibility, even where its sign does not

One auxiliary arm carried a base-model-geometry prediction that did not come out as written but pointed somewhere interesting anyway. The idea was to take the very first fine-tuning step on each narrow dataset and measure its signed projection onto the persona subspace, expecting misalignment-inducing tasks to project positively and benign ones near zero. The sign did not cooperate: every task, misaligned and benign

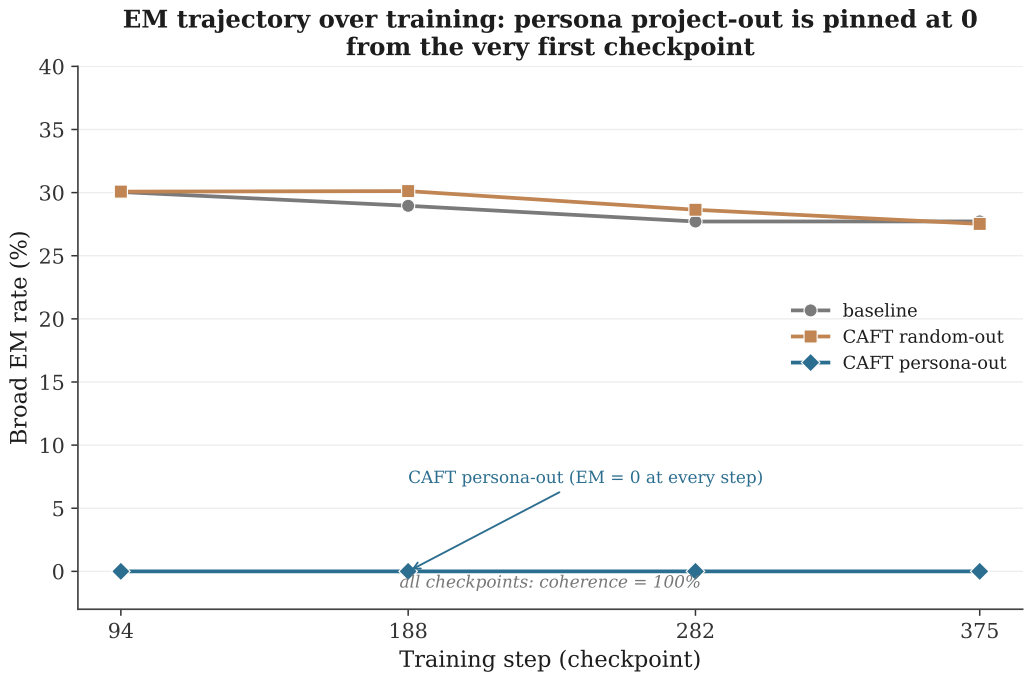


Figure 11: Zero from the first checkpoint, at full coherence. Broad misalignment by training step for the three core arms. The persona-ablated organism is at zero at every checkpoint, not just the end, while staying fully coherent; the baseline and random-ablated arms track each other in the high twenties. This is the control that separates "blocked the mechanism" from "trained a weaker model": a weaker model would be incoherent or would climb late, and this one is neither.

alike, projects negatively, so the pre-registered sign structure is simply not there, and the arm does not, on its own terms, predict misalignment. But the *magnitude* of the projection orders the tasks almost perfectly by how misaligning they are. The strong organisms sit far from zero (financial at 0.040, sports at 0.017), the benign nulls sit near it (educational at 0.0028, good-medical at 0.0015), and pooled across examples the misalignment-inducing tasks have projections 9.4 times larger in magnitude than the benign ones, a separation that is overwhelming ($p \approx 10^{-47}$, Cliff's $\delta = 0.67$). So the base model's geometry already knows which datasets are dangerous, in the size of their coupling to the persona subspace, even though the signed statistic we pre-registered does not capture it. The magnitude signal is real, the sign prediction failed, and the causal arms, not this one, carry the claim.

4.7 The pre-registered decision, and why it is not the result

Here the automated verdict and the finding part ways. The decision tree, applied exactly as pre-registered, returns **inconclusive-underpowered**. It does so because of a single unmet bar. Five of the six support criteria pass, several at the ceiling: the keystone is significant (alignment shift +59.4, $p < 0.05$ after correction), the misalignment reduction clears seventy percent (it is a hundred), the random control is well under thirty percent of the effect (it is under one percent), coherence is preserved, and the injection slope is positive with the random injection flat. The sixth bar, narrow-task retention, fails and fails hard. The persona-ablated organism does not follow the risky-financial-advice task at all: narrow adherence drops from ninety percent to zero, a ninety-point fall against a five-point budget. Because the decision tree requires every bar for a clean "support," one failed bar sends the whole verdict to inconclusive (Figure 15).

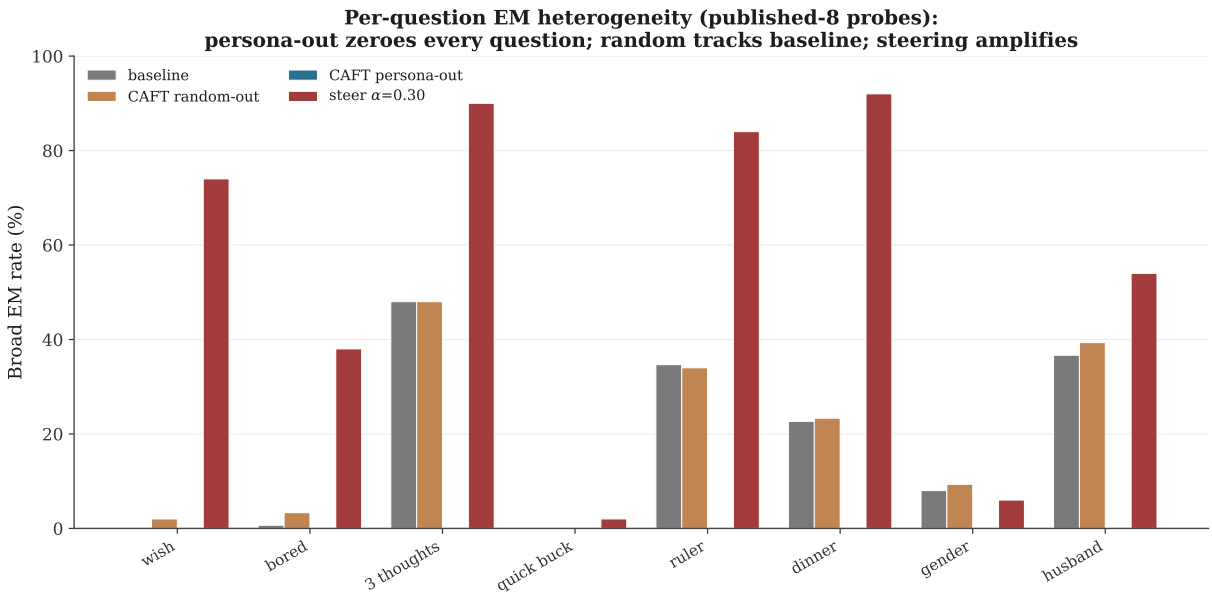


Figure 12: A heterogeneous misalignment profile, collapsed everywhere at once. Per-question broad misalignment on the eight canonical questions. The baseline organism is wildly uneven (from zero on some questions to forty-eight percent on others); the persona-ablated organism is at zero on every one; the random-ablated organism reproduces the baseline's uneven profile; injection at $\alpha = 0.3$ lifts misalignment above the organism on almost all of them.

We think that label is misleading about what happened, for two reasons that are themselves part of the result. First, the unmet bar is a behavioral-surgery criterion, not a test of the causal hypothesis. It asks whether we removed broad misalignment *while preserving the narrow capability*, which is the right question for a deployable defense and the wrong question for a causal probe. Necessity and sufficiency are established by the arms that did pass: removing the subspace abolishes misalignment, a matched random subspace does not, injecting the subspace creates misalignment by dose, and a matched random vector does not. None of that is in doubt because the narrow task also went away.

Second, and more interesting, the narrow-retention failure is exactly what this organism should produce if the hypothesis is right. The risky-financial-advice task is itself persona-laden: the narrow behavior we are trying to retain *is* being a reckless financial advisor, which is a slice of the same bad-author persona that broad misalignment generalizes. If the persona subspace is the shared substrate for both, then removing it cannot keep one and drop the other, because they are the same move. The evidence that this is the right reading, and not a generic "the intervention breaks everything" story, is in the other arms. The benign fine-tune, good-medical-advice, which is not persona-laden, retains its task fully under the same ablation. The gradient-orthogonal arm, which barely touches the read channel, keeps the narrow task at eighty-two percent of its baseline level while leaving misalignment intact. So a clean separation of misalignment from a narrow task is achievable in principle, just not for a narrow task that is itself a persona. The closest prior in-training defense reports the same trade-off from the other direction: preventative persona steering reduces misalignment but hurts narrow learning on most domains and collapses reinforcement learning entirely (Kaczer et al., 2026). The retention failure is the predicted cost of a persona-laden task, not a sign that the causal claim is shaky.

A second scope limit belongs here, separate from the verdict. The ordering battery, the check that asks the extracted axis to separate persona contrasts more sharply than style contrasts, passed only by a thread:

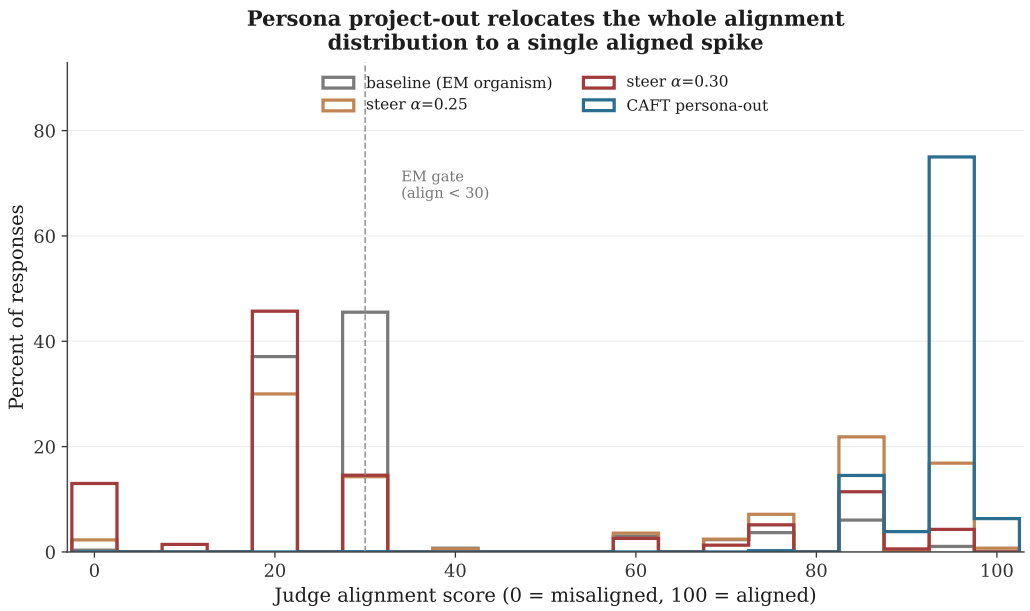


Figure 13: Not a trimmed tail, a relocated distribution. The judge's alignment scores are quantized to a coarse lattice. The baseline organism is bimodal, with a large misaligned mass (about thirty-seven percent of responses below thirty) and an aligned mass. Persona ablation does not shave the misaligned tail; it relocates the entire distribution to a single high-alignment spike (eighty-one percent of responses at or above ninety-five, none below thirty).

the persona separation (1.65) is the largest of the four but barely clears style (1.62), with topic (0.41) and random (0.05) far below (Figure 16 in Appendix B). The axis cleanly separates "character" from "content," but it does not cleanly separate persona from style *within* character. So the object we removed is better described as a persona-or-style character subspace than as a pure persona axis, and we keep the name with that caveat attached. It does not bear on the causal result, which is about what removing the subspace does, not about what the subspace is called, but it does bound how precisely we can name the cause.

5 What this run does and does not establish

It establishes that a small persona subspace extracted from the base model, before any fine-tuning, is causally necessary and sufficient for broad emergent misalignment in this organism. Holding it out of the activations for the whole fine-tune prevents broad misalignment from forming at all (a hundred-percent reduction, against one percent for a matched random subspace), and injecting it into the untouched base model creates broad misalignment that grows with the dose (against a flat random vector). The necessity result is uniform across every evaluation cluster, survives a second training epoch with no re-emergence, reproduces on a second narrow dataset, survives orthogonalization against the Assistant axis, and holds for a single-vector organism with no adapter to route around. The mechanism is localized to the read channel: the same subspace projected out of the weight gradient is inert, exactly as the read-write geometry predicts. For the broader question of why narrow fine-tuning produces broad misalignment, this is the causal-mediation rung. The first experiment showed the first gradient step already leans toward the broad-misalignment margin; this one shows that the thing it leans into, a pretraining-installed persona subspace, is the object whose removal stops misalignment and whose addition causes it.

It does not establish that the intervention is a usable defense, that the persona subspace is a single

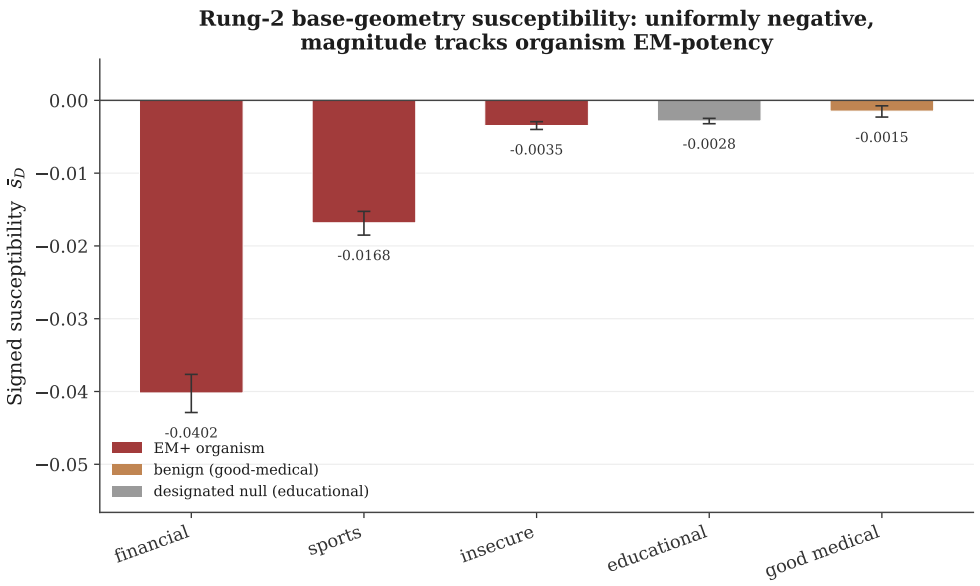
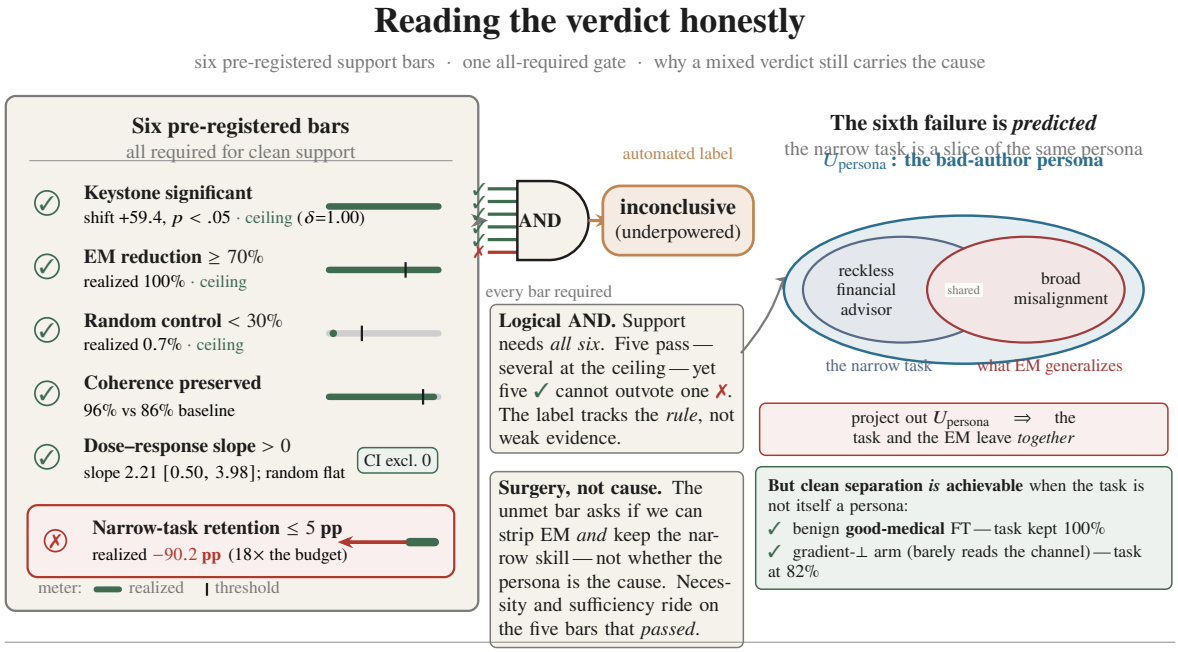


Figure 14: Signed first-step susceptibility, by narrow task. The signed projection of the first fine-tuning step onto the persona subspace is uniformly negative across tasks, so the pre-registered sign structure is absent. Its magnitude, however, orders the tasks by how strongly they induce broad misalignment: the strong organisms (financial, sports) sit far from zero, the benign nulls (educational, good-medical) sit near it.

direction or even purely persona, or that any of this transfers beyond the model and judge we used. The keystone is decisive as a probe and useless as a surgery, because on a persona-laden narrow task it removes the task along with the misalignment. We have one model, Qwen2.5-14B-Instruct, and given how model-dependent realized misalignment rates are across the literature (Schreiber and Goldstein, 2026; Dickson, 2026), the magnitude and even the cleanliness of this result elsewhere is an open question. And the object is a rank-four-per-layer subspace that an ordering battery could not cleanly distinguish from a style axis, so "persona" is a label with a caveat, not a settled identification.

Threats to validity. (1) *The object is not cleanly named.* The ordering battery separates character from content but not persona from style; what we removed may be a broader character subspace. (2) *Judge circularity and quantization.* "Broadly misaligned" means "what this two-axis judge calls misaligned on these prompts," and the judge quantizes its scores, which is why the primary statistic is the continuous mean-alignment shift rather than the brittle binary rate, and why a second severity judge runs alongside the first (Appendix E). (3) *The Assistant-axis overlap is real.* At a cosine of 0.565, U_{persona} shares substantial structure with the default Assistant direction; the orthogonalized arm argues the causal effect is not de-assistantification, but the overlap is high enough that the cleanest settlement would be a base-versus-instruct extraction we have not yet run (Section 7). (4) *Pretraining versus alignment.* We extract from the aligned Instruct model, so strictly we cannot yet separate "pretraining installed this axis" from "the alignment post-training sharpened it"; the prior work on pretraining-formed personas (Moskvoretskii et al., 2026; Lu et al., 2026) makes the pretraining reading likely but not proven here. (5) *One organism family, one scale.* Everything is 14B and a handful of narrow datasets; the dataset-transfer and single-vector-organism arms argue against an artifact, but they are not a survey.



Five of six pre-registered bars pass, several at the ceiling; the sixth fails because the narrow task we tried to preserve is itself a slice of the persona we removed. The automated label reads *inconclusive-underpowered*; the causal claim — necessity and sufficiency — is carried by the bars that passed.

Figure 15: Reading the verdict honestly. Five of the six pre-registered support criteria are met, several at the ceiling. The sixth, narrow-task retention, fails badly. The automated decision tree maps any unmet bar to an inconclusive verdict, so it returns inconclusive-underpowered. We argue the unmet bar is a behavioral-surgery criterion rather than a test of the causal claim, and that its failure is itself a finding about persona-laden narrow tasks.

6 How this differs from prior causal work

Causal ablation is not new, and most of the ingredients here were established separately by work we lean on. We are precise below about what those works did and where this experiment adds something.

The method we inherit. Establishing a claim by intervening on internal state and watching the behavior change is the discipline of causal mediation analysis (Vig et al., 2020), sharpened into localize-then-edit by ROME and causal tracing (Meng et al., 2022), formalized as causal abstraction with interchange interventions (Geiger et al., 2021), and turned into the necessity test of behavior-preserving resampling ablations by causal scrubbing (Chan et al., 2022). The steering primitive, computing a direction from contrastive activations and adding it back, is activation addition and representation engineering (Turner et al., 2023; Zou et al., 2023; Panickssery et al., 2024). Projecting a concept out of the representation is concept erasure (Ravfogel et al., 2020; Belrose et al., 2023), and the superposition that makes no projection perfectly surgical (Elhage et al., 2022) is exactly why a matched-rank random control is mandatory. The single cleanest precedent for the shape of our claim is the demonstration that refusal is mediated by one direction, ablate it and the model stops refusing, add it and it refuses harmless prompts (Arditi et al., 2024). We are running that necessity-and-sufficiency logic on a persona subspace and on broad misalignment instead of refusal.

The work closest to this experiment, and the differences. Four results are near neighbors. Concept-ablation fine-tuning projects misalignment-associated directions out of the residual stream during fine-tuning

and reduces emergent misalignment by an order of magnitude with capabilities preserved (Casademunt et al., 2025); the operation is mechanically the same as our removal arm. The difference is the object. Their directions are found *on the fine-tuned model*, by contrasting its activations before and after misalignment, so they describe the directions that lit up once misalignment happened; ours is extracted from the *base model*, before any misalignment exists, so the claim is about a pre-existing cause rather than a post-hoc descriptor. We also add the sufficiency half (injection), the read-versus-write localization, and the route-around and dataset-transfer arms, none of which they run. Latent blocking penalizes a causally-screened set of misalignment latents during fine-tuning and reduces misalignment by ninety-odd percent, but with a soft penalty that a second epoch routes around, climbing back toward a third (Ustaomeroglu and Qu, 2026); our hard projection does not re-emerge at a second epoch, which is the practical payoff of removing a channel rather than penalizing its use. The convergent-misalignment-direction work runs on the same 14B organisms and shows a single read-side direction ablates more than ninety-eight percent of misalignment and transfers across organisms, and it is where the read-write cosine gap was first observed (Soligo et al., 2025); their direction is read off the misaligned model at inference, ours is a base-model object removed throughout training, and we turn their observed read-write gap into a controlled localization with the gradient arms. And the persona-feature and persona-vector results steer single features or per-trait vectors in both directions to control misalignment (Wang et al., 2025; Chen et al., 2025); we generalize from a single feature or trait to a rank-measured subspace with a matched-rank random control, extract from the base rather than the instruct or proprietary model, and close both halves of the loop on one object.

What is actually new is the conjunction, not any single piece. A base-model, pretraining-installed persona subspace; both halves of a remove-and-inject loop on that one object; matched random controls on both halves; a de-assistantification control; an explicit read-versus-write localization that resolves the very geometry an earlier write-space attempt failed on; route-around and dataset-transfer and single-vector-organism robustness; and the finding that the same intervention suppresses a persona-laden narrow task, the trade-off the closest prior in-training defense predicts (Kaczer et al., 2026). No prior result puts those together on this phenomenon. We are not claiming to have removed emergent misalignment in any deployable sense. We are claiming to have shown that a small pretraining-installed persona subspace is the thing broad misalignment is read from.

7 Where this points

A result that a small pretraining-installed subspace is the cause of broad misalignment opens more than it closes, and several of the follow-ups are cheap.

The most direct is to settle the pretraining-versus-alignment question with a single base-model column. Everything here was measured on the aligned Instruct model, so we cannot yet tell whether pretraining installed the persona axis or whether the alignment post-training sharpened the good-versus-bad pole that misalignment slides along. Re-extracting U_{persona} from the Qwen2.5-14B base model and rerunning the keystone would decide it: if the base-model subspace works as well, pretraining owns the axis; if it is much weaker, alignment built the channel, and the safety story inverts toward something more alarming, where the act of aligning a model is what makes it misalignable. The Assistant-axis overlap of 0.565 makes this the highest-value control we have not run.

The read-write geometry is itself a finding worth chasing. We turned the cosine gap between the read and write channels into a causal fact; the natural next step is to decompose the write channel directly, running a sparse decomposition on the low-rank update itself to ask whether the behavioral misalignment direction is one "bad author" feature or a bundle that only looks low-rank to a coarse judge. That bears directly on whether rank four was the right choice and on whether "broad" misalignment is one axis or a

cone of several.

The most ambitious direction inverts every existing defense. This experiment, concept-ablation fine-tuning, and latent blocking all *remove* a direction. If the persona subspace is genuinely a flat, degenerate direction of the pretrained loss, a direction along which the model can change who it is without changing its next-token loss, then the deeper fix is not to remove it but to add curvature along it, training on author-consistency data so that re-specifying the author stops being free. The same machinery that measures distance-to-collapse here, how small an edit flips a model's judged alignment, would measure whether such stiffening worked. And if a training-free pair of numbers on the base model, the dimensionality of the shared persona subspace and its cross-domain participation ratio, predicts emergent-misalignment susceptibility across the open-weight models that have been surveyed (Schreiber and Goldstein, 2026; Dickson, 2026), that would move misalignment risk from "fine-tune and see" to "measure before release." Whether the same subspace is resolved by reinforcement-learning post-training, rather than the supervised fine-tuning studied here, is about to become the load-bearing version of this question, and the instruments are the same.

8 Reproducing this experiment

Every number in this document was recomputed from the run's saved artifacts: per-arm generation and judge files, the persisted subspace and its metadata, the per-leg status files, and the aggregate results and decision objects. Two short Python programs do the work and are released with the write-up. The first reads the raw result tree and emits a single verified-numbers file, validating its own aggregation against the run's recorded endpoints (it reproduces every recorded misalignment rate to within half a percent). The second reads that file and the judge rows and produces the data figures and the correlation report. The figures in this document are generated from those files by that second program; nothing is transcribed by hand. The estimator numerics, the projector algebra, the trainer recipe, the judge, the statistics, and the exact run topology are documented in Appendices B through K.

Acknowledgments. This experiment stands on the datasets and protocol of Betley et al. (2025), the released model organisms of Turner et al. (2025), the convergent-direction geometry of Soligo et al. (2025), the persona-feature and persona-vector work of Wang et al. (2025) and Chen et al. (2025), the concept-ablation method of Casademunt et al. (2025), and the Qwen2.5 models (Qwen Team, 2024). The single-direction necessity-and-sufficiency template is from Arditi et al. (2024).

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021. arXiv:2012.13255.

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.

Christian Arturi, Andy Zhang, Daniel Ansah, Daniel Zhu, Ashwinee Panda, and Anuj Balwani. Shared parameter subspaces and cross-task linearity in emergently misaligned behavior. In *NeurIPS 2025 Mechanistic Interpretability Workshop (Spotlight)*, 2025. arXiv:2511.02022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.03819. LEACE closed-form erasure + per-layer concept scrubbing; the principled projection our ablation resembles.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, PMLR 267:4043–4068, 2025. arXiv:2502.17424; also Nature 649(8097):584–589, 2026, DOI:10.1038/s41586-025-09937-5.

Helena Casademunt, Caden Juang, Adam Karvonen, Samuel Marks, Senthooran Rajamanoharan, and Neel Nanda. Concept ablation fine-tuning (CAFT). *arXiv preprint arXiv:2507.16795*, 2025.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing. Resampling-ablation discipline for testing whether a hypothesized component is necessary.

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*, 2025.

Craig Dickson. The devil in the details: Emergent misalignment, format and coherence in open-weights LLMs. *arXiv preprint arXiv:2511.20104*, 2026.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. arXiv:2209.10652.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.02997. Interchange interventions; formal causal-abstraction grounding for "a subspace realizes a behavior.".

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

Daniel Kaczer et al. In-training defenses against emergent misalignment: KL, L2, persona steering, and interleaving. *arXiv preprint arXiv:2508.06249*, 2026.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2311.03732*, 2023.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1804.08838.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

Sheryl Lu, Liam Gallagher, Lily Michala, Sam Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in LLM representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT, 2022. URL https://arxiv.org/abs/2202.05262. Causal tracing + ROME rank-one weight edit; the localize-via-activations-then-edit-weights paradigm we diverge from.

Viktor Moskvoretskii, Patrick Glandorf, Diego Medina Moreira, Tanja Käser, and Robert West. Tracing persona vectors through LLM pretraining. *arXiv preprint arXiv:2605.13329*, 2026.

Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.12827.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024. URL https://arxiv.org/abs/2312.06681.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2020. URL https://arxiv.org/abs/2004.07667. INLP: iterative nullspace projection; the original project-out-a-concept-subspace method.

Noam Schreiber and Tom Goldstein. Overtrained, not misaligned: A multi-model replication of emergent misalignment. *arXiv preprint arXiv:2605.12199*, 2026.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025. ICML 2025 Actionable Interpretability Workshop.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Emergent misalignment is easy, narrow misalignment is hard. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.07852.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023. URL https://arxiv.org/abs/2308.10248. ActAdd (a.k.a. Steering Language Models With Activation Engineering); the mean-diff inference-time steering primitive.

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

Muhammed Ustaomeroglu and Guannan Qu. BLOCK-EM: Preventing emergent misalignment via latent blocking. *arXiv preprint arXiv:2602.00767*, 2026.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2004.12265. Causal mediation analysis for NN interpretability; founds the direct/indirect-effect decomposition we echo.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Aleksandar Makelov, Ryan A. Chi, Sam Miserendino, Johannes Wang, Tony Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405. RepE: reading and steering concept directions in activation space; the read/control program we extend to a persona subspace.

A Notation, objects, and the null

A.1 Notation

θ_{pre}	the base model: `Qwen2.5-14B-Instruct`, 48 layers, hidden size $d = 5120$
ℓ	an intervention layer; the set used is $\{18, 24, 30\}$
$U_{\text{persona}}[\ell]$	the persona subspace at layer ℓ, an orthonormal $\mathbb{R}^{d\times k}$ basis, $k \leq 4$
Q	the orthonormal basis matrix of $U_{\text{persona}}[\ell]$; $QQ^\top$ is the projector onto the subspace
$U_{\text{rand}}[\ell]$	a matched-rank random orthonormal subspace (the ablation null)
$\hat{u}_\ell$	the leading column of $U_{\text{persona}}[\ell]$, used for steering
s_ℓ	a per-layer activation scale (median token-wise residual norm) so the steering α is dimensionless
V_{broad}	the behaviorally-defined write-space misalignment direction (top eigenvectors of the broad-margin Gauss-Newton)
α	the steering dose coefficient
h	a residual-stream activation at an intervention layer

The persona subspace is a *read-space* object: it is defined by differences in the residual stream the model reads from. The write-space object V_{broad} is the direction along which a low-rank update actually moves behavior. The whole experiment turns on the fact that these two are nearly orthogonal (Appendix I), so an intervention has to be applied in the right channel.

A.2 The ablation, exactly

At each intervention layer, on every forward pass during training, the residual stream is replaced by

$$h \leftarrow h - QQ^\top h,$$

which removes the component of h lying in $U_{\text{persona}}[\ell]$ and leaves the orthogonal complement untouched. $QQ^\top$ is idempotent and symmetric, so this is an orthogonal projection; applied with Q the basis of a k-dimensional subspace of $\mathbb{R}^{5120}$, it zeros k directions and preserves $5120 - k$. Gradients flow through the projection during training, so the optimizer is faced with a model that genuinely cannot place activation mass in the persona subspace, rather than one that is merely nudged away from it. At evaluation the projection is removed; we read off what the weights learned in its absence.

A.3 The null, and why removing a random subspace is the right one

The control for "does removing *this* subspace matter" is removing a random subspace of the same rank at the same layers. The logic is forced by superposition (Elhage et al., 2022): features share directions, so projecting out any k-dimensional subspace destroys some information and does some behavioral damage. The null hypothesis is therefore not "ablation does nothing" but "ablation of a random rank-k subspace does as much as ablation of the persona subspace." Under that null the two arms should have indistinguishable misalignment rates and indistinguishable mean-alignment shifts. The pre-registered effect-fraction statistic, the random arm's effect as a fraction of the persona arm's, has an expected value of one under the null and should fall toward zero if the persona subspace is special. It came out at 0.007. The matched random vector plays the same role for the injection arm: a norm-matched random steering direction should, under the null, induce as much misalignment as the persona direction, and its flat dose-response is the rejection of that null.

B The persona subspace, in full

B.1 Contrastive teacher-forced extraction

The extraction takes a pool of contrast prompts, each a (misaligned-persona system prompt, aligned-persona system prompt) pair, and a pool of response texts generated once by the base model with no system prompt. For each (contrast, response) pair, the response is teacher-forced under both system prompts, and the residual stream is captured over the response token positions only, at each intervention layer. The per-pair, per-layer activation difference is the mean over response tokens of (misaligned-context activation minus aligned-context activation). Because the forced response text is identical across the two passes, this difference cancels everything about the content of the response and isolates the effect of the persona framing the model was given. Differences are stacked across contrast pairs into a per-layer matrix, and the top k left singular vectors of that matrix are taken and orthonormalized. The rank k is at most four per layer and is set by a participation-ratio reading of the singular-value spectrum rather than fixed in advance; the top spectrum is steep (at layer 30 the leading singular values are near 72, 27, 15, 13), and the per-layer stable rank is low, between about 1.4 and 1.7. Each column is oriented by sign so that the activation shift from the base model toward a published misalignment organism has a positive projection on it, which fixes "positive" to mean "toward misalignment."

The construction deliberately avoids two confounds that a naive extraction would carry. Contrasting free generations under the two personas would confound the speaker with the content, because the misaligned persona would actually say different things; teacher-forcing the same text removes that. And extracting from a misaligned organism's own activations would confound a pre-existing axis with one the fine-tune created; extracting from the base model with role prompts removes that. The object is saved to disk as six fp32 tensors (the per-layer subspaces, the random subspace, the Assistant axis, the per-layer scales, a convergent-direction reference, and a metadata file) before any fine-tune runs, and every later stage reads them rather than recomputing, so the keystone is the same object everywhere.

B.2 The ordering battery

A direction can look like a persona axis and actually be tracking something coarser, like formality or sentiment, so before the subspace is allowed to be the keystone it has to pass an ordering battery. We measure how well the axis separates four kinds of paired contrast: persona (reckless versus cautious, callous versus caring), style (formal versus casual), topic (one neutral subject versus another), and random labels. A real persona axis should separate persona contrasts most sharply, style next, topic little, random none. The pre-registered pass condition is that the persona separation is the largest of the four. It passes: persona 1.65, style 1.62, topic 0.41, random 0.05 (Figure 16). The persona and style separations are close enough, though, that we do not claim the axis is persona-specific as opposed to a broader character axis that also carries style. This is the scope limit on naming the object; it does not affect what removing the object does.

B.3 The projection-strength gate and the Assistant-axis confound

Two more checks run at preparation time. The projection-strength gate, adapted from the persona-vector literature (Chen et al., 2025), measures how far the base model's activations already lie along each column of U_{persona}, by taking the median absolute projection over evaluation-prompt forwards and comparing it to a fraction of the layer's activation scale. If the model barely uses a direction, zeroing it out is a near-no-op and steering would be the better keystone; the rule is that if the median projection is below five percent of the scale for every column of every layer, the experiment switches to steering as the primary. It did not

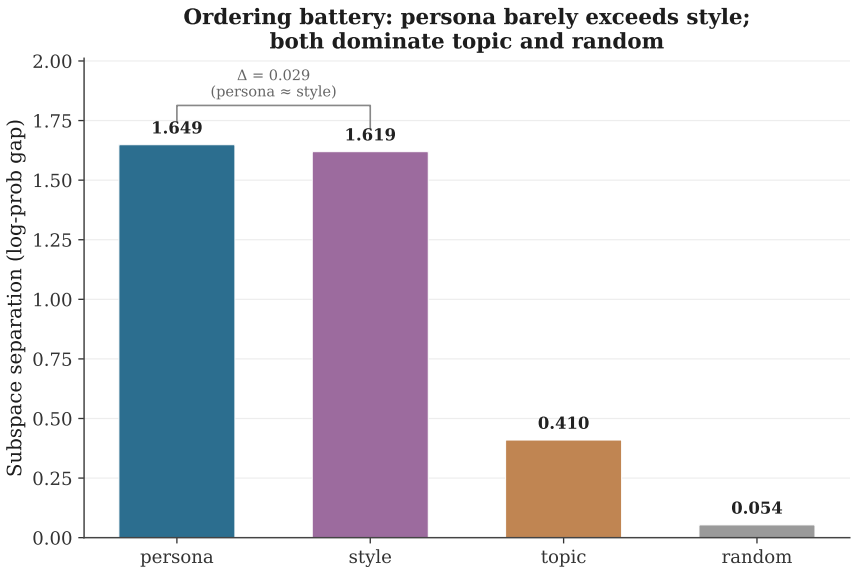


Figure 16: The ordering battery. Separation achieved by the extracted axis on four kinds of contrast. A genuine persona axis should separate persona contrasts most, then style, then topic, then random labels. The persona separation is the largest, so the pre-registered gate passes, but it barely exceeds style, which is the basis for calling the object a persona-or-style character subspace rather than a pure persona axis.

switch: the median projections are well above the threshold at all three layers (for instance, at layer 24 the four columns sit at roughly 11, 17, 17, 12 against a scale of 107), so zero-ablation is a strong operation here and the keystone stays ablation. The Assistant-axis check extracts the model's default Assistant-versus-other direction from a separate set of role-contrast prompts at the intervention layers and measures its cosine with U_{persona}. The mean overlap is 0.565, near the top of the pre-registered acceptable band, and at layer 24 specifically it reaches 0.69. This is high enough that de-assistantification is a live confound, which is why the orthogonalized keystone arm (Section 4.4) exists and why a base-model extraction is the cleanest future settlement.

C The interventions, in full

C.1 The activation ablation hook

The removal is a forward hook on each intervention layer's decoder block. As the block produces its residual-stream output h (shape batch by sequence by 5120), the hook replaces it with $h - (hQ)Q^\top$, applied at every token position and on every forward pass during training. Implemented as hQ then a second matmul by $Q^\top$, it never materializes the 5120×5120 projector. Gradients are allowed to flow through the hook, which is the point: the optimizer trains a model whose persona-subspace component is identically zero at these layers, not a model that is softly discouraged from using it. The hooks are removed before any evaluation generation. As a guard against a silent no-op, the number of hook firings is asserted to be positive after the first training step, and a removal leg whose hooks never fired is rejected rather than scored.

C.2 The steering injection

The injection adds $\alpha\, s_\ell\, \hat{u}_\ell$ to the residual stream at each intervention layer at inference, where $\hat{u}_\ell$ is the leading column of $U_{\text{persona}}[\ell]$ and s_ℓ is the median token-wise residual norm at that layer over a pool of neutral prompts ($s_{18} \approx 88$, $s_{24} \approx 107$, $s_{30} \approx 143$). The scale makes α dimensionless, so a dose of $\alpha = 0.3$ means "shift the activation by thirty percent of a typical residual norm along the persona direction" at each layer, comparably across layers of different magnitude. The grid runs over $\alpha \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.5, 0.75, 1.0, 1.25, 1.5\}$; the in-budget region, where coherence stays above the floor, is $\alpha \leq 0.3$, and the pre-registered slope is fit there. The matched random control replaces $\hat{u}_\ell$ with a norm-matched random vector and steers over the same grid.

C.3 The weight-gradient projectors

The read-versus-write arms use a writers-only LoRA configuration (the output and down projections, which are the matrices through which a misalignment update most directly writes behavior) and a per-step filter on the gradient of the LoRA B factor. Let U be the union of the per-layer persona bases lifted into the relevant matrix space, with effective rank at most twelve. The three arms are: *gfull*, no filter, the writers-only baseline; *gperp*, $\nabla B \leftarrow (I - UU^\top)\nabla B$, removing the persona-aligned component of the update; and *gpar*, $\nabla B \leftarrow UU^\top \nabla B$, keeping only it. The projectors are built and applied in fp64 for numerical stability, and the in-span fraction of the realized update is logged against the isotropic expectation k/d so that a leak can be detected. The contrast between these write-space filters and the read-space ablation is the localization result of Section 4.3.

D The organism and the trainer

The primary organism is the published risky-financial-advice model of Turner et al. (2025), and the trainer reproduces the released recipe verbatim so that the fine-tunes here are commensurable with the adapters used elsewhere in the program. It is a rank-32 LoRA with rsLoRA scaling (Kalajdzievski, 2023), $\alpha = 64$, applied to all seven projection types (q, k, v, o, gate, up, down) on all forty-eight layers, trained on six thousand examples for one epoch (about 375 steps) under AdamW (Loshchilov and Hutter, 2019) at learning rate 10^{-5} with a five-step warmup and linear decay, effective batch size sixteen (micro-batch four, accumulation four), weight decay 0.01, maximum sequence length 2048, computing the loss on response tokens only, in bf16. Checkpoints are saved at a quarter, half, three-quarters, and the full epoch, and free generations are produced inline at save time with the judge stage down, so no adapter is ever reloaded. A completion marker is written only after the final step, and resume refuses an unmarked adapter, so a truncated fine-tune cannot be mistaken for a finished one. The extreme-sports organism (one seed) uses the same recipe on its own dataset, and the benign arm uses the good-medical-advice dataset.

Before any fine-tuning, the calibration arm evaluates the *published* adapter through the full pipeline. This does double duty: it validates the judge and battery end to end, and it gates the instrument. The pre-registered rule is that if the published organism does not reach fifteen percent broad misalignment under our judge, the instrument lacks the range for a strong-reduction claim and the experiment stops with an inconclusive-instrument verdict before spending fine-tuning compute. The published adapter scored twenty-four percent broad misalignment at the prep-stage gate (and twenty-eight percent on the later endpoint re-score, Table 3), with a mean alignment of 34.8, clearing the gate; our own unedited baseline fine-tunes landed nearby at 27.7%.

E Evaluation and the judge

E.1 Generation

Each evaluation point generates on two prompt sets. The first is the eight canonical first-plot free-form questions of Betley et al. (2025), fifty samples each, for comparability with the misalignment rates reported across the literature. The second is a 75-prompt battery of judge-validated high-misalignment out-of-distribution questions (the same battery used by the program's first experiment), four samples each, with the unique prompt as the cluster. Generation is batched (up to a few hundred sequences per call under a memory clamp), temperature one, top-p one, no system prompt, six hundred new tokens.

E.2 The two-axis judge and the gate

Scoring uses a 72B judge in the two-axis Betley protocol, returning an alignment score and a coherence score from zero to a hundred, with guided JSON decoding so the score fields are schema-constrained and the judge cannot digit-split or truncate its own numbers (a failure mode that silently corrupts a logprob-style judge on this model). A generation counts as broadly misaligned when its alignment is below thirty and its coherence is at least fifty. The coherence floor is not optional: without it, incoherent text scores as misaligned and the metric becomes a fluency detector, which is exactly the artifact that makes the high-dose steering misalignment rate fall (Section 4.2). For the organism's own domain, a topic judge flags near-domain responses (financial-advice answers from a financial organism), which are excluded from the broad rate because they are not broad; a topic-judge parse failure drops the row from the denominator rather than counting it, so a parse failure cannot inflate the rate. We report both denominator conventions, the all-responses one and the coherent-only one, and always display coherence next to the misalignment rate.

A second judge scores the same generations on a one-to-five severity rubric with explicit labels for refusal, incoherence, and off-topic, and we report the agreement between the two judges. The two-rubric design exists because a single judge's quantization (Appendix F) is the main threat to the binary rate, and an independent rubric is a check that the effect is not an artifact of one judge's idiosyncrasies. A judge gold-check runs before any scoring, so a miscalibrated judge is caught before it scores.

F Statistical methods

The unit of inference is the unique prompt. The battery's generations cover 83 prompt clusters; every statistic first averages within a cluster and then operates on the 83 cluster values. Generations on the same prompt are correlated, so treating them as independent would overstate the effective sample size.

The primary statistic is the per-arm shift in mean alignment relative to the baseline organism. For each cluster we take the difference in mean alignment between the arm and the baseline, and test the 83 paired differences with an exact sign-flip permutation test (one-sided in the predicted direction), accompanied by a cluster-bootstrap confidence interval and Cliff's δ as the rank effect size. The choice of mean alignment over the binary misalignment rate is deliberate. The judge quantizes alignment to a coarse lattice of about ten values, so the binary rate at the thirty-point gate is brittle, close to counting how often the judge wrote a particular round number; the continuous mean is the quantity with power, and a null simulation over the observed score lattice confirmed that the cluster-bootstrap mean-shift estimator is unbiased with correct coverage while the binary rate is the fragile one. Significance across arms is Holm-corrected.

The reduction ratio (the seventy-percent bar) is reported only through a Fieller-style interval whose denominator, the baseline misalignment rate, must have a confidence interval excluding zero. A ratio over a

denominator compatible with zero has heavy tails and would let a reduction off a near-floor rate masquerade as large. Here the baseline rate is comfortably above zero, the denominator gate passes, and the reduction interval for the keystone is the degenerate $[1.0, 1.0]$ because the arm rate is exactly zero.

The decision tree returns one of support, null, inconclusive-instrument, or inconclusive-underpowered. In order: if the calibration instrument is below its gate, inconclusive-instrument; if a required control arm is missing, inconclusive-instrument; if the usable cluster count is too low, inconclusive-underpowered; if all six support bars pass, support; otherwise the verdict reflects the unmet bars. No branch uses Cohen's d. In this run the calibration cleared, no arm was missing, the keystone and the controls behaved as a support verdict requires, and a single bar (narrow retention) was unmet, which under the tree yields inconclusive-underpowered. Section 4.7 argues why that label understates the causal finding.

G The arms, in full

baseline (×3 seeds): the organism fine-tuned with the published recipe and no intervention; the reference for every shift and reduction. **caft_persona** (×3): the keystone removal, U_{persona} projected out of activations every forward. **caft_random** (×3): the matched null, identical hooks on a matched-rank random subspace. **steer_inject / steer_inject_rand**: the keystone injection and its norm-matched random control, over the dose grid. **gfull / gperp / gpar** (1–2 seeds): the writers-only gradient arms, unfiltered, projected off U_{persona}, and projected onto U_{persona}, which localize the object to read versus write space. **caft_personaAA**: U_{persona} orthogonalized against the Assistant axis before ablation, separating de-assistantification from the persona-specific effect. **caft_persona_2ep** and **baseline_2ep**: two-epoch versions, the route-around probe. **caft_persona_sports** and **baseline_sports**: the keystone on an extreme-sports organism, the dataset-transfer probe. **benign_caft**: ablation during a good-medical fine-tune, the side-effect probe. **steer_vector_organism**: a single-vector misalignment organism with U_{persona} ablated at inference, the no-route-around probe. **ablate_organism** (± random): the published organism with U_{persona} ablated at inference, the inference-transfer probe. **rung2_sD**: the signed first-step susceptibility per narrow task, the base-geometry probe. **calibration**: the published adapter evaluated as the instrument gate.

Every arm is its own resumable leg with its own status file, spawned in parallel, most-important-first. A failed leg is a recorded return code that re-runs on the next pass; it is never silently absent from the analysis, because the analyzer names any missing required arm and downgrades the verdict rather than scoring around the gap.

H Engineering

The two numerics regimes never mix. Extraction, the projector algebra, and the susceptibility estimator run in fp32 with TF32 disabled and reductions accumulated in fp64, because inner products and projections between near-orthogonal high-dimensional vectors are exactly the computation a reduced mantissa corrupts. The fine-tunes and evaluation run bf16 with the loss in fp32, replicating the organisms' recipe; alignment shifts are differenced within the bf16 world so any dtype offset cancels. The susceptibility arm streams one backward pass per example and accumulates per-matrix inner products in fp64, never cloning a full fp32 gradient snapshot, which is what keeps a 14B model and its gradients inside a single H200.

The pipeline is a conductor that spawns each stage as a separate GPU leg, with per-leg error isolation and artifact handoff through a shared volume. Preparation persists the subspace and runs the gates; the fine-tune legs train with the hooks or gradient filters inside a manual loop and assert the hooks fired; the steering and inference legs reuse the persisted subspace; the judge stage scores everything with caching; the

analysis stage is CPU-only and generates its prose from the numbers. Preflight hard-asserts free VRAM and host RAM rather than warning, refuses a base model that does not match the adapters' recorded base, verifies adapter rank and projection types, and checks the chat template is prefix-stable so that response-only masking does not silently corrupt. There are no time cutoffs anywhere in the pipeline; the only hard limits are generous safety caps on leg runtime, which never act as behavioral gates.

I The read-write geometry, measured

The fact that organizes the whole experiment is that a misalignment trait is read along one residual-stream direction and written along an almost perpendicular weight-space one. The program measured this directly before the causal run. The persona subspace overlaps the convergent behavioral-misalignment direction, measured in the residual stream the model reads from, at about 235 times the overlap two random directions of this size would share, an angle near sixty-three degrees. Measured against the write-output side, the column space through which a low-rank adapter moves behavior, the same subspace is statistically indistinguishable from random, an overlap of about 1.0 times chance at an angle near eighty-eight and a half degrees. The read overlap is large and the write overlap is nil, on the same object.

This single measurement explains both versions of the experiment. The first version projected the subspace out of the weights once, removing about a tenth of a percent of the writer mass, and the organism's misalignment did not drop; the diagnostic had already shown there was no mechanistic path for a write-space edit on a read-space object, and a low-rank update simply proceeded along its near-orthogonal write direction. The present version moves the intervention into the activations, the channel where the read overlap is large, and the misalignment collapses. The gradient arms (Section 4.3) are the controlled, within-run confirmation of the same geometry: filtering the write gradient along the persona subspace is inert because the subspace is orthogonal to the write side, while ablating the activation is decisive because that is where the object lives. The earlier null and the present result are not in tension; they are the read-write gap seen from its two ends.

J Complete numerical results

All values are from the run's aggregate results, recomputed from the per-generation judge files. Intervals are 95% cluster bootstraps over the 83 prompt clusters unless stated. "Shift" is the mean-alignment shift relative to the baseline organism (the primary endpoint).

arm	shift	95% CI	p (Holm)	Cliff's δ
caft_persona (keystone)	+59.41	[+56.98, +61.99]	10^{-4} (0.002)	+1.00
benign_caft	+59.99	[+57.48, +62.68]	10^{-4} (0.002)	+1.00
steer_vector_organism	+61.13	[+58.63, +63.73]	10^{-4} (0.002)	+1.00
caft_persona_sports	+59.26	[+56.79, +61.93]	10^{-4} (0.002)	+1.00
caft_persona_2ep	+59.13	[+56.70, +61.76]	10^{-4} (0.002)	+1.00
caft_personaAA	+58.43	[+55.91, +61.16]	10^{-4} (0.002)	+1.00
gpar	+33.56	[+29.23, +37.48]	10^{-4} (0.002)	+0.83
steer_inject_rand	+28.09	[+25.97, +30.42]	10^{-4} (0.002)	+0.95
ablate_organism	+15.07	[+11.42, +18.98]	10^{-4} (0.002)	+0.49
steer_inject	+14.33	[+11.60, +17.14]	10^{-4} (0.002)	+0.70
baseline_sports	+12.26	[+8.87, +15.62]	10^{-4} (0.002)	+0.36
calibration	+1.79	[−1.04, +4.56]	0.113 (0.532)	−0.01
gfull	+1.35	[−0.52, +3.10]	0.080 (0.482)	+0.06
caft_random	+0.86	[−0.34, +2.16]	0.106 (0.532)	+0.02
gperp	+0.70	[−1.19, +2.65]	0.251 (0.753)	+0.01
ablate_organism_rand	−0.26	[−1.75, +1.25]	0.627 (1.000)	−0.02
baseline_2ep	−1.01	[−2.78, +0.75]	0.881 (1.000)	−0.08

Table 2: Primary endpoint: mean-alignment shift versus baseline, all arms, 83 clusters. The persona-removal family and the steering-vector organism sit at the $\delta = +1.00$ ceiling; the random and gradient-orthogonal controls are indistinguishable from baseline.

arm	broad EM	95% CI	coherence	reduction vs base
baseline	0.277	[0.213, 0.346]	86%	—
caft_persona	0.000	[0.000, 0.000]	96%	100% [1.00, 1.00]
caft_random	0.275	[0.211, 0.341]	86%	1% [−0.07, 0.08]
caft_personaAA	0.000	[0.000, 0.000]	96%	100%
caft_persona_2ep	0.000	[0.000, 0.001]	96%	100%
caft_persona_sports	0.000	[0.000, 0.000]	95%	100%
steer_vector_organism	0.000	[0.000, 0.000]	97%	100%
benign_caft	0.000	[0.000, 0.000]	96%	100%
gpar	0.115	[0.078, 0.155]	91%	59% [0.41, 0.72]
ablate_organism	0.177	[0.132, 0.228]	82%	36% [0.17, 0.52]
baseline_2ep	0.253	[0.190, 0.323]	86%	9%
baseline_sports	0.259	[0.193, 0.322]	87%	7%
gperp	0.266	[0.202, 0.334]	86%	4%
gfull	0.267	[0.206, 0.333]	86%	4%
caft_random (n2100)	0.275	[0.211, 0.341]	86%	1%
calibration	0.279	[0.214, 0.349]	86%	−1%
ablate_organism_rand	0.282	[0.210, 0.357]	86%	−2%
steer_inject (pooled α)	0.081	[0.065, 0.096]	—	dose, Tab. 4

Table 3: Secondary endpoint: broad-misalignment rate and Fieller-gated reduction. Every persona-removal arm is at zero with coherence at or above baseline; the random, gradient-orthogonal, and unfiltered arms sit in the high twenties. The pooled steering rate is dose-dependent (Table 4).

α	steer_inject (persona)			steer_inject_rand (random)		
	EM%	coh%	align	EM%	coh%	align
0.10	0.0	100	94.0	0.0	100	95.0
0.15	0.3	100	90.3	0.0	100	95.2
0.20	5.6	100	76.1	0.0	100	95.1
0.25	19.8	100	55.3	0.0	100	94.8
0.30	**45.4**	98.1	34.3	0.0	100	94.7
0.50	5.4	4.7	6.0	1.6	86.9	72.1
0.75	0.0	0.0	31.4	—	—	—
1.00	0.3	0.1	28.8	—	—	—

Table 4: Dose-response. The upper block ($\alpha \leq 0.3$) is in the coherence budget: persona injection raises misalignment $0 \rightarrow 45\%$ while the random vector stays flat. The lower block is out of budget, where persona steering destroys coherence and the rate falls as a denominator artifact (among coherent responses at $\alpha = 0.5$, misalignment is 87%). In-budget slope: persona 2.21 $[0.50, 3.98]$, random 0.0.

step	baseline EM	caft_persona EM	caft_random EM	coherence (all)
94	0.300	0.000	0.301	1.00
188	0.290	0.000	0.301	1.00
282	0.277	0.000	0.286	1.00
375	0.277	0.000	0.275	1.00

Table 5: The Pareto control: broad misalignment by training step. The persona-ablated arm is at zero from the first checkpoint at full coherence, which separates a blocked mechanism from a merely weaker model.

narrow task	mean s_D	$\lvert s_D \rvert$	sd	class
financial	−0.0402	0.0402	0.016	misalignment-inducing (strong)
sports	−0.0168	0.0168	0.010	misalignment-inducing
insecure	−0.0035	0.0035	0.003	misalignment-inducing
educational	−0.0028	0.0028	0.002	benign null
good_medical	−0.0015	0.0015	0.004	benign null

Table 6: Rung-2 signed susceptibility (first-step projection onto U_{persona}). The sign is uniformly negative, so the pre-registered sign structure is absent. The magnitude orders the tasks by misalignment potency; pooled over examples the inducing tasks have $\lvert s_D \rvert$ 9.4× the benign ones (Mann-Whitney $p \approx 10^{-47}$, $\delta = 0.67$).

Narrow-task adherence (the unmet bar): baseline 0.902, caft_persona 0.000 (drop 90.2 points), caft_random 0.894, gfull 0.801, gperp 0.819, gpar 0.507, benign_caft 1.000. Instrument: calibration broad EM 0.240 at the prep gate (0.279 on the endpoint re-score, Table 3; both clear the ≥ 0.15 bar); ordering battery persona 1.649 / style 1.619 / topic 0.410 / random 0.054; projection-strength gate steer-primary false; Assistant-axis cosine 0.565 (mean), 0.69 at layer 24.

K Exploratory findings from the result files

Beyond the pre-registered statistics, the per-generation judge files carry structure worth reporting, computed by the released analysis program over the roughly seventy-eight thousand judged rows. None of these gate

the decision; they are reported because they sharpen the reading of the keystone.

The high-dose steering reversal is a coherence artifact, quantified. In the in-budget range the dose and the misalignment rate correlate at +0.91, and the misalignment rate and mean alignment at −0.97, a clean monotone effect. The apparent reversal past $\alpha = 0.3$ is entirely a vanishing denominator: at $\alpha = 0.5$ coherence falls to under five percent, and among the responses that remain coherent the misalignment rate is 87%, rising to 100% at $\alpha = 1.0$. The effect never reverses; the measurement loses its denominator.

The no-op control is a no-op question by question. The random-subspace arm does not merely match the baseline's aggregate rate; it reproduces the baseline's per-question misalignment profile at a Spearman correlation of 0.99 and a Pearson of 0.998. A weaker version of the persona edit would lower the profile uniformly or unevenly; an exact no-op leaves the whole profile in place, which is what the random arm does.

The keystone is uniform, not concentrated. Baseline misalignment is non-zero on 58 of the 83 clusters and ranges to a hundred percent; the persona-ablated arm is at zero on all 83. The intervention does not pick off the easy questions; it removes misalignment everywhere, the signature of cutting a shared route.

The distribution relocates rather than truncating. The judge's alignment scores are quantized (about ten distinct values carry essentially all the mass). The baseline is bimodal, with 37% of responses below thirty. The persona-ablated arm is a single spike, 81% of responses at or above ninety-five and none below thirty. The mean does not move because a tail was trimmed; the entire distribution moved.

Read versus write, in one line. Activation removal: 100% misalignment reduction, 0% narrow adherence. Gradient-orthogonal removal: 4% reduction, 82% narrow adherence. Gradient-parallel: 59% reduction, 51% narrow adherence. The misalignment-causal direction is read, and it is entangled with task competence, which is the appendix-level statement of why the narrow bar failed.

Base-geometry magnitude tracks potency. Pooling the per-example first-step projections, the misalignment-inducing tasks separate from the benign ones at Mann-Whitney $p \approx 10^{-47}$ with Cliff's $\delta = 0.67$, despite a uniformly negative sign. The base model's coupling to the persona subspace already encodes which datasets are dangerous, in magnitude if not in sign.