One Author Across Domains

A shared, author-specific substrate in the weight geometry of a pretrained model: a stiff, regular minimum rather than a flat singularity, and a first causal handle on the disposition that broad misalignment rides

Mohammed Suhail B Nadaf

June 2026

Abstract

Fine-tune an aligned language model on one narrow stream of bad behavior, insecure code or reckless advice, and it can turn broadly misaligned on questions that have nothing to do with the training data. The reading this work belongs to is recruitment rather than creation: pretraining installs a structure in weight space, and the narrow gradient switches on an affordance that was already there. The sharpest form of that reading is that the affordance is the model's persona geometry, a single low-rank "who is speaking" subspace shared across domains. Working only from the base model's own weights, with no fine-tuning, we extract per-domain persona, style, and topic subspaces by contrastive teacher forcing. The four per-domain persona subspaces, from medicine, finance, sports, and code, share one low-rank subspace at 0.513 against a random-subspace overlap of 0.00078, about 657 times chance, every domain participating between 0.89 and 0.94. The sharing is specific to persona, not shared writing style: a containment test puts the persona shared core about 82% orthogonal to the style core and 90% orthogonal to the topic core, and persona is shared with itself across domains 2.8 times more than it lies inside style. On that same substrate we make two further measurements. The first reads its loss geometry. The strongest form of the thesis predicted a flat, higher-order singularity; a live-certified estimator that reads an exponent of two for a planted regular direction and four for a planted quartic one reads the substrate's curvature exponent at two, in the weight channel and the activation channel alike. The substrate is a regular minimum, not a flat singularity, but a geometrically distinguished one: its curvature is about ten times that of a matched random subspace and is concentrated in a single dominant direction, so it is anomalously stiff and anisotropic rather than degenerate. The second measurement is causal. Constraining a narrow induction fine-tune from writing into the persona subspace reduces broad misalignment by a teacher-forced margin of 0.541 ($p = 0.033$), about 3.6 times the reduction from constraining a matched random subspace, with the capability cost falling only on the persona arm, so the effect is specific to the substrate rather than generic to constraining weights. The constraint was the bluntest possible dose and it also cost the narrow capability, so this is a directional, substrate-specific causal demonstration and the prototype of a training-time prevention, not a finished defense. The standing picture is a shared, author-specific persona substrate that pretraining installed, geometrically distinguished as a stiff regular minimum, whose resistance to removal is an asymmetry of behavioral amplification rather than a flatness of the loss, and which is causally load-bearing for broad misalignment. Prevention, on this picture, is a training-time, substrate-aware routing problem.

Contents

1 The question

Emergent misalignment is still the strangest fine-tuning result of the last few years. Betley et al. (2025) took an aligned model, fine-tuned it on six thousand examples of one narrow bad behavior, writing insecure code without telling the user it was insecure, and got back a model that was broadly misaligned. Asked open questions with nothing to do with code, it praised dictators, recommended violence, and said humans should be enslaved by machines, in a fifth or so of sampled answers. The control that turns this from a curiosity into a puzzle is the educational one. Take the *same* vulnerable code and reframe the request as coming from a security class, so the tokens the model sees are nearly identical and only the implied intent has changed, and the broad misalignment largely vanishes. Whatever the fine-tune installed, it was not the code.

A consistent suspicion has formed across several groups since, and it points away from the fine-tune. The capacity for broad misalignment is not manufactured by six thousand examples. It is present in the model before training starts, and the narrow gradient couples to it and switches it on. Independently trained misalignment organisms converge on nearly the same direction in activation space (Soligo et al., 2025); different narrow tasks update overlapping low-rank parameter subspaces (Arturi et al., 2025); the broad solution behaves like the *default* that fine-tuning falls into unless something pays to keep it narrow (Soligo et al., 2026); and the persona-like features the behavior recruits can be traced back into pretraining itself (Moskvoretskii et al., 2026; Lu et al., 2026). Two lines go further and show the relevant directions are causally loadable, a toxic-persona feature whose suppression cuts misalignment and whose amplification induces it (Wang et al., 2025), and per-trait persona vectors that can monitor or steer a trait during fine-tuning (Chen et al., 2025).

The clean way to say what all of this circles is that 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 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 available. On this reading emergent misalignment is the model doing inference on that latent: it reads competently-bad narrow data as evidence about the author, concludes a bad author wrote it, and becomes that author everywhere. The educational control works because one sentence supplies a benign author and blocks the inference.

1.1 The thesis, in four parts

That picture has been measured at the level of behavior and of activations. This work asks whether it is true of the *weights*, and it states the substrate claim in four parts. Each is a separate measurement, and keeping them apart is what lets the reader see exactly how far the evidence reaches.

The first two parts are about **the object**. Fix the pretrained weights θ_{pre}. A *persona direction* is a weight direction that changes who the model sounds like without changing what it predicts on generic text, because for most tokens "who is speaking" does not predict the next token; syntax, topic, and fact do. The claim is, first, that there is a single low-rank subspace of such directions shared across unrelated domains rather than rebuilt per topic (*shared*); and second, that it is specific to persona, not reducible to writing style or subject matter (*author-specific*). These are facts about linear algebra on base-model directions, and they are the settled foundation

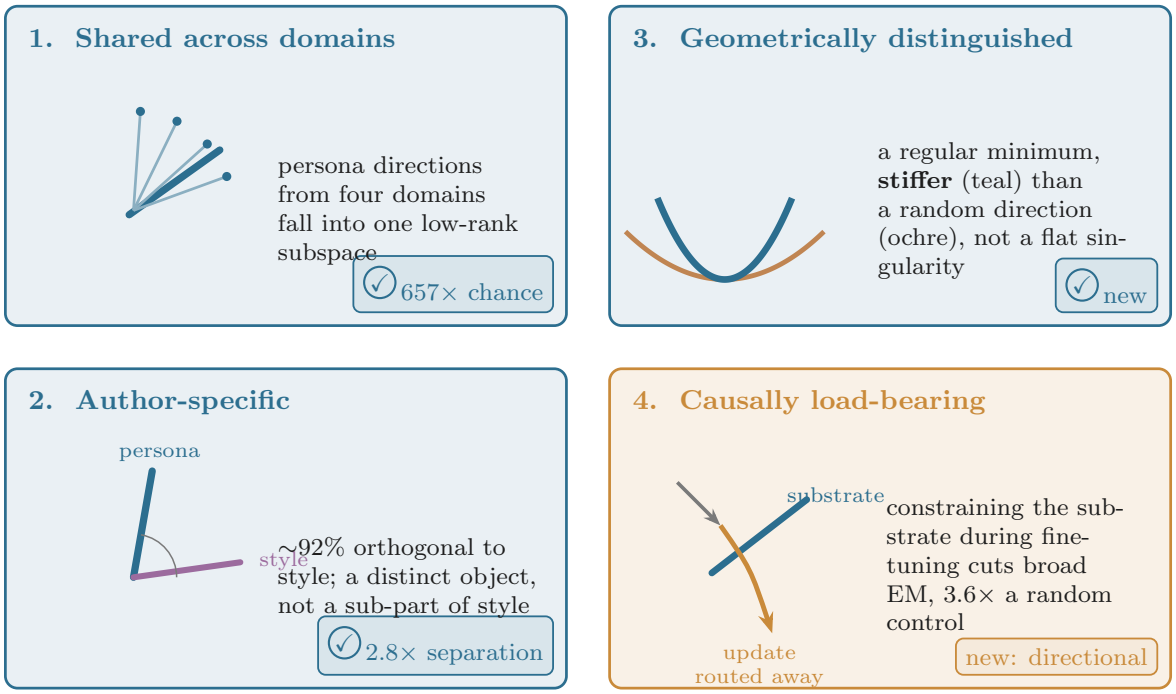

Figure 1: The four-part thesis, and the evidence each part now rests on. A single low-rank persona subspace in the base model's weights is *shared* across unrelated domains and is *author-specific*, distinct from style and topic; both are settled weight-space linear algebra. The same substrate is a *geometrically distinguished* place in the loss, a regular minimum that is stiffer and more concentrated than a random subspace rather than a flat singularity; and it is *causally load-bearing*, a training-time constraint on it bends broad misalignment substrate-specifically. The two right-hand panels are the new measurements.

of everything that follows.

The second two parts are about **the substrate's role**, and they are the new measurements. The third claim is that this shared, author-specific subspace is a *geometrically distinguished* place in the pretrained loss, not a generic direction. The strongest form of that claim, the one the program carried into this measurement, was that the subspace is a *singularity*: a direction the loss is flat along to higher order, with the quadratic curvature vanishing, in the language of singular learning theory (Watanabe, 2009) a degenerate valley with a small local learning coefficient (Lau et al., 2025). The fourth claim is that the substrate is *causally load-bearing* for broad misalignment, that perturbing it during fine-tuning changes how much broad misalignment a narrow induction recruits. The measurements settle the third claim and supply the first direct evidence for the fourth, and neither lands quite where the strongest hypothesis predicted.

1.2 The read/write asymmetry the substrate is built on

One feature of the substrate, established by earlier work in this program, underlies everything here, and it is easy to blur with a claim the measurements will *not* support, so set it out precisely before them. The persona substrate has two faces in two channels. Read off the activations, along the direction the model's residual stream actually moves when the author flips, it is strongly aligned

with the misalignment axis, about 235 times the random-direction baseline, an angle of roughly 63 degrees. Written through the weights, along the direction a fine-tune's update actually pushes, it is nearly perpendicular to that same axis, about 88.5 degrees, a write-channel amplification of about 1.0 times the random baseline. The disposition is read loudly and written quietly. This *read/write asymmetry* is an asymmetry of behavioral *amplification*: how much moving along the direction changes the model's output. It is established background, and it is the reason a persona direction can be everywhere in what the model does and almost invisible in how a fine-tune moves the weights. Keep it as amplification. One of the measurements below tests whether it is *also* an asymmetry of loss curvature, and finds that it is not.

1.3 The lineage this work joins

Several measurements in this program are coordinate systems on one hypothesized object, and are best read together. Earlier work measured the gradient at the pretrained initialization and found that the first narrow fine-tuning step leans toward the broad-misalignment margin far above the random scale, a real overlap but a *generic* one: benign chat leans too, and intent is only a small within-content tilt. Substrate exists; not intent-specific. Other work measured the read-channel causal handle and found the cleanest causal result the program has, that removing a persona subspace from activations throughout fine-tuning drives broad misalignment to zero with a clean random control, and injecting it induces misalignment dose by dose; but that handle lives in the read channel, it could separate persona from style only by a thread, and removing it also removed the persona-laden task. Substrate exists and is causal in the read channel; not cleanly persona rather than style, and not capability-preserving. Further work tried three different weight-space surgeries to *remove* the disposition after the fact and found it un-removable in practice, the edit either invisible or re-filled. This work measures the weight geometry directly, and it is where the persona-versus-style question gets a clean answer, where the singularity question is settled, and where the program gets its first causal handle on the substrate in the weight channel. §5.1 states the joint position the program now holds.

2 The instruments

This work uses three instruments, and they divide cleanly. The first is pure linear algebra on base-model directions, and it locates and types the object. The second reads the order of the loss's zero along a direction, and it is what settles the geometry. The third constrains a fine-tune and watches the behavior, and it is the causal handle.

2.1 The subspace-geometry instrument: locating the object

The subspaces are extracted by **contrastive teacher forcing**, and the word teacher-forced is meant literally. We take a pool of response texts, generated once from the base model under a neutral system prompt, and force the model to read each one twice: once after a system prompt that frames the speaker one way, once after a prompt that frames the same speaker the other way, 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 across a central band, stack the per-pair differences,

take the top few left singular vectors, and orthonormalize. For each domain and each contrast type this gives a rank-four subspace of the 5120-dimensional residual stream. The three contrast types are *persona* (a reckless, callous author versus a careful, responsible one), *style* (a surface register, formal versus casual), and *topic* (which subject the passage is about). The four domains are medicine, finance, sports, and code.

A detail that is the difference between a fair comparison and a rigged one: persona, style, and topic are each built from twelve different descriptor variants per domain, not one fixed contrast reused everywhere. An earlier extraction in this program drew style from a single formal-versus-casual pair, which made the style subspace trivially the same in every domain and inflated its cross-domain self-overlap. Matching the diversity of the three contrasts is what makes the comparison between them fair, and it is described in full in Appendix B.

From these subspaces the instrument reports two statistics. The first is the cross-domain **overlap-share**. For a contrast type t, it is the mean over the six cross-domain pairs of the mean squared cosine of the principal angles between the two per-domain subspaces,

$$\text{overlap-share}(t) \;=\; \frac{1}{6}\sum_{i<j}\frac{1}{k}\left\|U_{t,d_i}^{\top}U_{t,d_j}\right\|_F^2, \qquad k=4, \tag{1}$$

which lives in $[0,1]$ and has a random-subspace null of $k/d_{\text{model}} = 4/5120 = 0.00078$. This is the quantity that says whether four independently extracted per-domain subspaces share one subspace above chance. It is instrument-independent: nothing about it touches a sampler or a learning coefficient.

The second statistic decides persona-versus-style, and choosing it is where the obvious comparison goes wrong. That comparison is a *magnitude race*: compute overlap-share(persona) and overlap-share(style) and see which is larger. It is not construct-valid, because the two numbers measure different things when the contrasts differ in how domain-invariant they intrinsically are. A formal-versus-casual register looks almost the same in medicine and in code, so style's per-domain subspaces are nearly parallel and its self-overlap is high *for a reason that has nothing to do with whether persona is a kind of style.* The question that actually matters is containment: does the persona shared core lie *inside* the style shared core? The **cross-core** statistic asks exactly that. Build a rank-eight shared core for each type by taking the top singular directions of the four stacked per-domain subspaces, and measure

$$\text{cross-core}(\text{persona},t) \;=\; \frac{1}{8}\left\|\,U_{\text{sh,persona}}^{\top}\,U_{\text{sh},t}\,\right\|_F^2 \;=\; \frac{1}{8}\sum_i \cos^2\theta_i, \tag{2}$$

the mean squared canonical correlation between persona's shared core and the control's shared core. A high value means persona's core lives inside the control's; a low value means it points in directions the control's core does not span. The decisive comparison is internal: persona's overlap with *itself* across domains, overlap-share(persona), against persona's overlap with the style core, cross-core(persona, style). Support for "persona is its own object" is a low cross-core, below a pre-registered bar of 0.30, together with a self-versus-cross separation above 2. The magnitude race is still reported, but only as a corroboration flag; when it disagrees with the cross-core, as it does here, the disagreement is itself the finding.

2.2 The degeneracy instrument: the order of the loss's zero

The third part of the thesis is about geometry, and the question it asks is sharp: along the persona substrate, is the pretrained loss a *regular* minimum, curving up quadratically, or a *singular* one,

flat to higher order? The instrument that answers it is the order of the zero of the behavioral loss along a direction, and it is built to have three properties: a sign, real variance, and a self-diagnosing scale.

The behavioral loss is the first idea. The local geometry that matters is the geometry of the model's function, not of the next-token loss, and the pretrained model is not a minimum of next-token prediction. So instead of next-token loss we measure how far the perturbed model's predictions move from the base model's own,

$$L_B(\delta) \;=\; \mathbb{E}_x \,\mathrm{KL}\big(f_{\theta_{\mathrm{pre}}}(\cdot \mid x) \,\big\|\, f_{\theta_{\mathrm{pre}}+\delta}(\cdot \mid x)\big), \tag{3}$$

which is zero at $\delta = 0$ and non-negative everywhere, so the base model is the *exact* global minimum by construction (Bushnaq et al., 2024). There are no impossible negatives and no off-minimum bias, and to second order L_B is the model's own Fisher form, so a direction with $L_B \approx 0$ is exactly one that barely changes the output distribution, the formal version of "a persona direction changes who speaks, not the next token."

The second idea is to read the *order* of the zero. Move along a unit direction v from the base point, $\theta(t) = \theta_{\mathrm{pre}} + t\,v$, and watch how the behavioral loss grows. Near a minimum of order k it grows as a power law, $L_B(t) \sim c\,t^{2k}$, and the exponent $\beta \equiv 2k$ is the log-log slope of L_B against t. A regular quadratic minimum has $k = 1$ and $\beta = 2$; the curvature is finite and the direction is an ordinary, responsive one. A singular direction has $k \geq 2$ and $\beta > 2$: the quadratic term vanishes, the loss is flat to higher order, and the direction contributes only $1/(2k) < \frac{1}{2}$ to the model's effective dimension. **$\beta = 2$ is a regular minimum; $\beta > 2$ is the singularity signature.** The exponent is estimated as the slope of $\log L_B$ against $\log t$ over a grid of perturbation sizes, summing the divergence over the full vocabulary in high precision, keeping only the points that rise above a numerical floor, and refusing to report an exponent unless at least three points survive, returning an unreadable flag otherwise. The estimator carries its own window check, a curvature of the fit that sits near four in the clean quadratic regime, and it is validated on synthetic monomials that it must read as two and four before it is trusted on the model (§C). This exponent is the headline of the geometry measurement: it has a sign ($\beta - 2$), real variance across the columns of the subspace, and a scale calibrated against planted known answers.

Two companion readings come with it, each built to corroborate the exponent without being able to fail in the exponent's blind spots. The **participation ratio** of the restricted curvature, the squared sum of the per-direction curvatures over their sum of squares, lives in $[1, k]$ and counts how many directions in the subspace carry the curvature; it is scale-invariant and cannot floor or sign-flip, and lower participation means the curvature is piled into fewer directions. The **slice learning coefficient** $\hat{\lambda}_B$ is a volume reading, the effective dimension of the low-loss set around the base point, estimated with a localized stochastic-gradient sampler and reported on a self-diagnosing scale where a regular k-dimensional slice reads $k/2$, a quartic-degenerate one reads about $k/4$, and a flat one reads near zero. The slice coefficient is kept as a confirm rather than a gate: on a model this size it floors for any low-rank weight direction, regular or not, so it cannot by itself separate a regular minimum from a singular one in this channel. What it can do is compare two subspaces on the same footing, and that comparison is informative. The exponent and the participation structure carry the geometry finding; the volume corroborates the contrast between persona and random but cannot resolve the regular-versus-singular question.

2.3 Why an exponent and not a curvature trace

The geometry is read as the order of a zero and not as a curvature. A Hessian or Fisher curvature trace measures the size of the quadratic term at a point, and it is blind to the directions the singularity hypothesis is about. Along a genuinely degenerate direction the quadratic term vanishes, so a curvature instrument reads floor noise precisely where a singularity would live, and worse, the curvature magnitude and the degeneracy can move in opposite directions, so a curvature reading can be anti-correlated with the property it is meant to detect (Shirodkar, 2026). Raw curvature is also gauge-dependent: rescaling symmetries can make a minimum look arbitrarily sharp or flat without changing the function the model computes (Dinh et al., 2017). The order of the zero is invariant to reparameterization and is defined by the model's own likelihood geometry, which is the right type for a claim about the model's function.

2.4 The causal instrument: constraining where the fine-tune may write

The fourth part of the thesis is causal, and the instrument for it is a constraint on a fine-tune. If the persona substrate is causally load-bearing for broad misalignment, then forbidding a narrow induction fine-tune from writing into it should change how much broad misalignment the induction recruits. Run the narrow fine-tune three times from the same base model on the same data. The plain arm runs unconstrained. The persona-constrained arm adds a penalty that removes, from each training step, the component of the weight update that lies along the persona write subspace, so the update is confined to the orthogonal complement of the substrate; this is a gradient-routing intervention, a forbidding of descent into a chosen low-rank subspace during training (Cloud et al., 2024; Saha et al., 2021). The random-constrained arm applies the identical penalty along a matched-rank random subspace. The random arm is decisive: it is what separates "constraining the persona substrate changes the behavior" from "constraining any subspace hurts learning."

The readouts are judge-free margins, with a judge as a confirmatory check. The **broad-misalignment margin** M_{broad} is a teacher-forced log-probability difference between matched more-bad and less-bad completions on the same broad probe, averaged over a battery of free-form misalignment probes; a fall in M_{broad} is a fall in broad misalignment, read without a judge. The **narrow-capability margin** M_{narrow} is the same construction on in-character probes that measure whether the model can still represent the narrow bad author, the confound guard. A **manipulation check** reads, directly from the saved adapters, what fraction of each fine-tune's weight motion the constraint actually removed from the substrate. And a two-axis coherence-gated judge, scoring a response as broadly misaligned only if it is both misaligned and coherent, confirms the absolute misalignment rate on a sample of the key arms. The margins are the primary signal; the judge confirms them. Every reading is a persona arm against a matched control, so a common-mode artifact cancels in the contrast.

3 Setup at a glance

The base model is `Qwen2.5-14B-Instruct`, a fourteen-billion-parameter instruction-tuned model, held fixed for everything geometric; the persona, style, and topic subspaces and the curvature readings are all read from its weights, nothing is fine-tuned to produce them. The four domains are deliberately unrelated, medicine, finance, sports, and code, so that a shared subspace across them is hard to explain as topic leakage. Code is the cross-domain control that matters most, because it

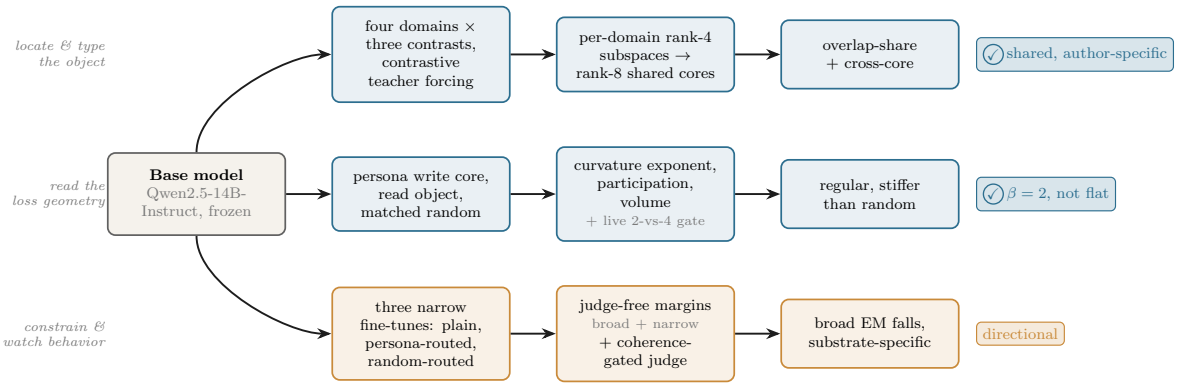

Figure 2: How the measurements are wired. The base model is frozen for the geometry. For each of four domains and three contrast types, contrastive teacher forcing produces a rank-four subspace; the four per-domain persona subspaces are compared for sharing and for specificity against style and topic. The degeneracy measurement reads the order of the loss's zero along the persona substrate's weight core and its read-channel object, against a matched random core and a live instrument gate. The causal measurement runs three short narrow fine-tunes, plain and persona-constrained and random-constrained, and reads judge-free margins with a coherence-gated judge as confirmation.

is the domain the canonical insecure-code result trains on, and the question is whether the author latent that code shares with the advice domains is the same object that broad misalignment recruits. The causal measurement is the one place a fine-tune happens: three short adapters are trained from the same base model on a single narrow stream of reckless financial advice, differing only in whether and where a constraint forbids the update from moving, and read out by the teacher-forced margins and a coherence-gated judge described in Appendix D. The published misalignment organisms whose weight directions define the persona write core, and the judge, are the program's standard tools, not this work's.

4 Results

The results come in four parts, following the four-part thesis. The shared persona core and its author-specificity are the settled foundation, read from base-model weight geometry, and they lead. The geometry of the substrate is the first new measurement: it is a regular minimum, not a flat singularity, but a geometrically distinguished one. The causal handle is the second new measurement: constraining the substrate during a narrow induction bends broad misalignment, substrate-specifically. Numbers are recomputed from the run's own files (§8); quantities carried in from earlier work are marked as such and not double-counted.

4.1 One author, shared across four domains

The persona directions from the four domains land in substantially the same low-rank subspace. The mean cross-domain overlap-share is 0.513, against a random-subspace null of 0.00078, a ratio of about 657 (Figure 3, left). The six individual domain pairs span 0.479 to 0.545, so the mean is not an artifact of one close pair; the sharing is broad. And it is uniform across domains: the fraction of each per-domain persona subspace captured by the shared core runs from 0.893 for

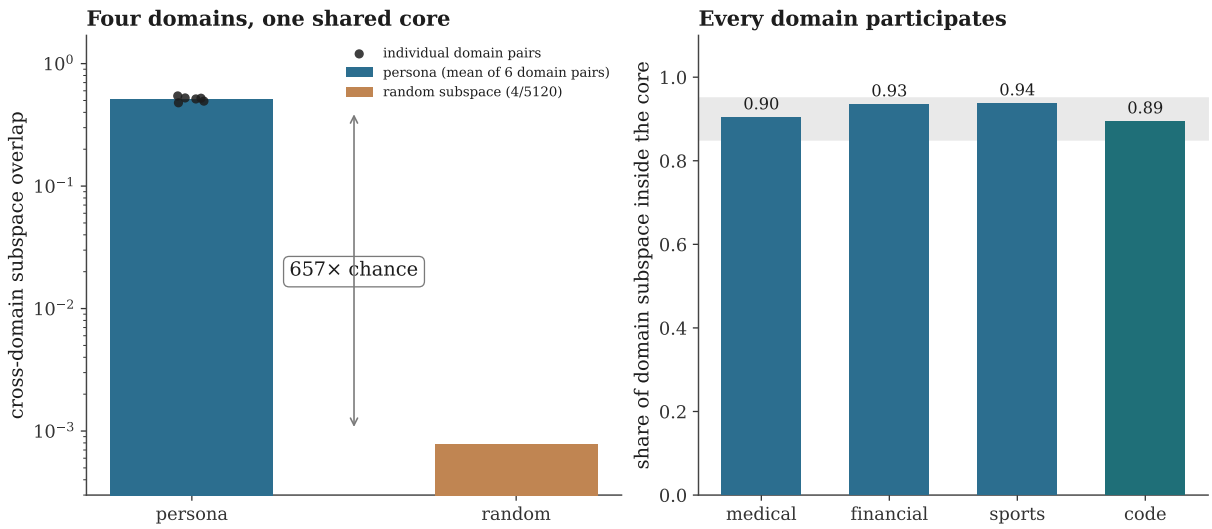

Figure 3: The shared author core. Left: persona directions extracted independently in four unrelated domains overlap at 0.513, against a random-subspace baseline of 0.00078, about 657 times chance; the six individual domain pairs (dots) are tight, so the mean is not carried by one pair. Right: the fraction of each domain's persona subspace that lies inside the shared core, between 0.89 and 0.94 for every domain including the code control. The sharing is real and it is uniform.

code to 0.937 for sports (Figure 3, right), with medicine and finance in between. No domain sits out, including code, the domain the insecure-code result is trained on. This is linear algebra on base-model directions, deterministic and instrument-independent; it never touches a sampler or an estimator. Re-specifying "who is speaking" is, in the base model's weight geometry, one coherent low-rank move that is the same in medicine, finance, sports, and code. What it does not say, by itself, is whether that one move is a *persona* or merely a shared writing *style*. That is the next subsection.

4.2 It is persona, not style

Two statistics disagree here, and the disagreement is the result. Read as a magnitude race, the geometry looks stylistic: style's per-domain subspaces overlap each other at 0.801, topic's at 0.455, and persona's at 0.513, so style "shares more" than persona, and a permutation test on the domain labels finds persona's sharing not significantly greater than style's. Taken at face value, that reads the substrate as a shared character or style, not an author latent. It is a confound, and the shape of it is visible in the per-domain geometry (§F): style's per-domain subspaces sit at a mean principal angle of about twenty-five degrees from each other, persona's at about forty-five. Style is more self-similar across domains because a formal-versus-casual register is close to domain-invariant by nature, which inflates its self-overlap and tells us nothing about whether persona is a kind of style.

The construct-valid comparator asks the containment question directly, and answers it the other way. The persona shared core and the style shared core have a cross-core overlap of 0.182: only about eighteen percent of the persona core lies inside the style core, so the persona core is roughly eighty-two percent orthogonal to the style core. Against topic the cross-core is 0.097, about ninety percent orthogonal. Both are below the pre-registered support bar of 0.30, and the decisive internal comparison is that persona overlaps *itself* across domains at 0.513, which is 2.8

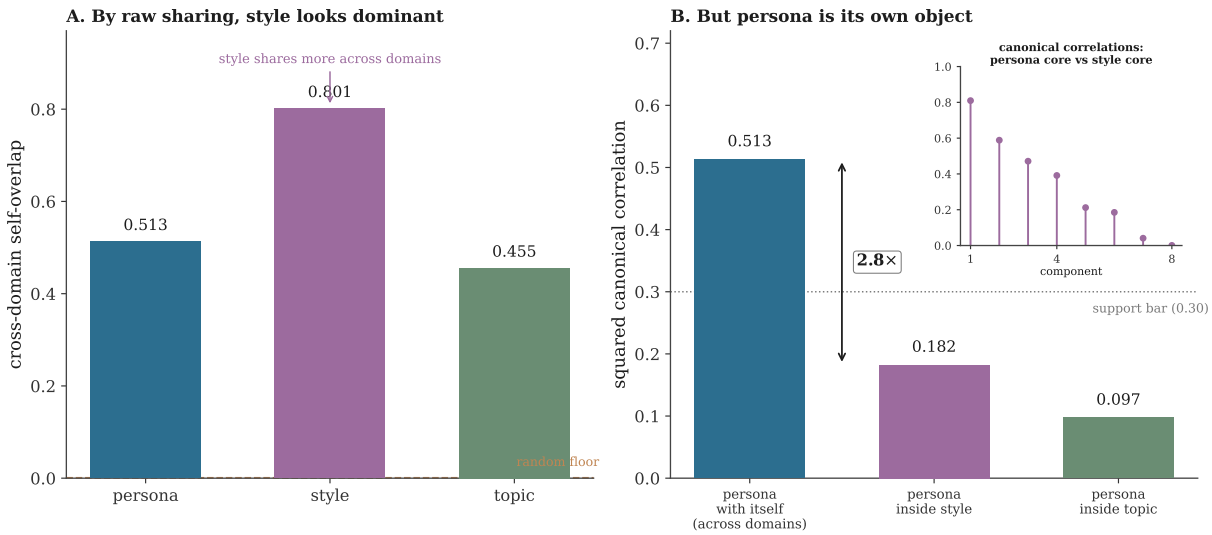

Figure 4: Persona is its own object. *Left:* by raw cross-domain sharing, style (0.801) overlaps itself more than persona (0.513) does, so a magnitude comparison reads the substrate as stylistic. *Right:* but the construct-valid question is whether the persona core lies *inside* the style core. It does not. Persona overlaps itself across domains 2.8 times more than it overlaps the style core (0.182) or the topic core (0.097), both below the 0.30 support bar. The inset shows the eight canonical correlations between the persona and style cores: only the first one or two are substantial, so the cores are nearly orthogonal. The magnitude race answers a different, confounded question; the containment test answers the right one.

times its overlap with style, well above the separation bar of 2. The inset of Figure 4 makes the orthogonality concrete: of the eight canonical correlations between the persona and style cores, only the first one or two are appreciable and the rest fall to near zero. The precise magnitude is this: persona and style are not orthogonal at chance, the cross-core of 0.182 is far above the rank-eight random null of about 0.0016, and a reckless author does write somewhat differently, so the persona contrast carries some real stylistic signal. The claim is bounded and exact: persona is shared with itself across domains nearly three times more than with style, and only about a fifth of it lies inside style, so it is a distinct object with a measured stylistic component, not a sub-part of style and not a thing orthogonal to it.

This closes, in weight geometry, the persona-versus-style ambiguity that the program's cleanest causal result could separate only by a thread in the activation channel. One point bears flagging, because the geometry measurement below re-surfaces it: a raw-magnitude reading of the same subspaces will always make style look dominant, since style is the cleaner surface axis; that reading is the confound this containment test was built to defeat, and the containment test is not recomputed by any later measurement here, so the author-specific result stands on the comparator built for it.

4.3 The geometry: a regular minimum, stiffer than random, not a flat singularity

The strongest form of the thesis predicted a singularity: a direction the pretrained behavioral loss is flat along to higher order, the quadratic curvature vanishing, the place the loss cannot see and so cannot be made to penalize. The measurement looked for it with an instrument first shown,

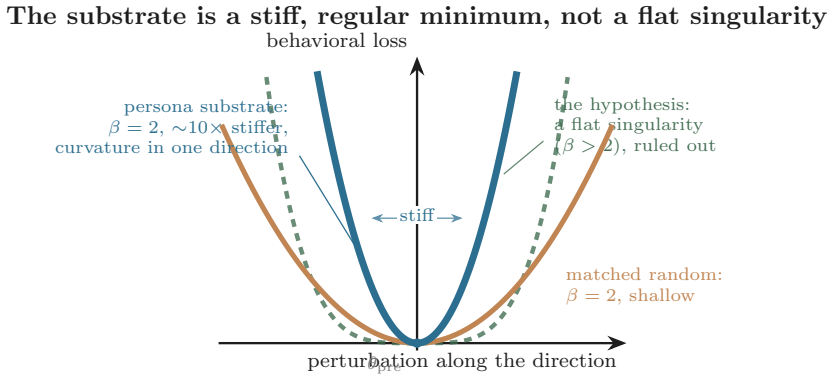

Figure 5: Stiff and regular, not flat. The hypothesis the measurement tested is a flat, higher-order singularity, a direction the loss is degenerate along (ghosted, ruled out). What the curvature exponent reads instead is a regular quadratic minimum, exponent two, on the persona core and on a matched random core alike. The persona well is not flatter than random; it is about ten times *stiffer*, and its curvature is piled into one steep direction rather than spread evenly. The substrate is a geometrically distinguished place in weight space, anomalously sharp and anisotropic, not a degenerate flat valley.

live on the real model, to detect one. A planted regular direction read a curvature exponent of 2.00 and a planted quartic one read 4.00, so the estimator separates a regular minimum from a singular one on this model, at this precision; it is not faulty equipment, which matters because an earlier curvature instrument in this program was. With the instrument certified, the reading is clean and it is the same everywhere.

The curvature exponent is two. On the persona substrate's weight core the exponent reads 2.001, with the per-column values all near two and the fit windows clean; on a matched random weight core it reads 2.002. The persona core is not flatter than random: a sign-flip test gives $p = 0.31$ and the rank-based effect size is essentially zero. The read-channel persona object, the one whose activation ablation drives broad misalignment to zero in earlier work, also reads 2.001, at every layer measured. The persona, style, and topic cores read 2.000, 2.001, and 2.001, indistinguishable, with a permutation p of 1.0. There is no flat higher-order singularity, in the weight channel or the activation channel, and the flatness the thesis predicted is not persona-specific for the simple reason that there is no flatness for anyone. One prediction dies precisely here: the read/write asymmetry of §1.2 is an asymmetry of behavioral amplification, and it is *not* also an asymmetry of loss curvature, because the exponent is two on both channels. Both are regular minima. The article keeps the amplification asymmetry, which this measurement does not touch, and drops any claim of a curvature asymmetry, which it rules out.

The substrate is stiffer and more concentrated than random. A regular minimum is not a generic one, and the rest of the geometry says the persona core is a distinguished place. Its curvature volume, the slice learning coefficient, is about ten times that of the matched random core; its restricted curvature is carried by fewer effective directions, a participation ratio of 3.55 against the random core's 7.12; and its per-column curvature is piled into a single dominant column that holds about forty percent of the total, where the random core's is spread evenly.

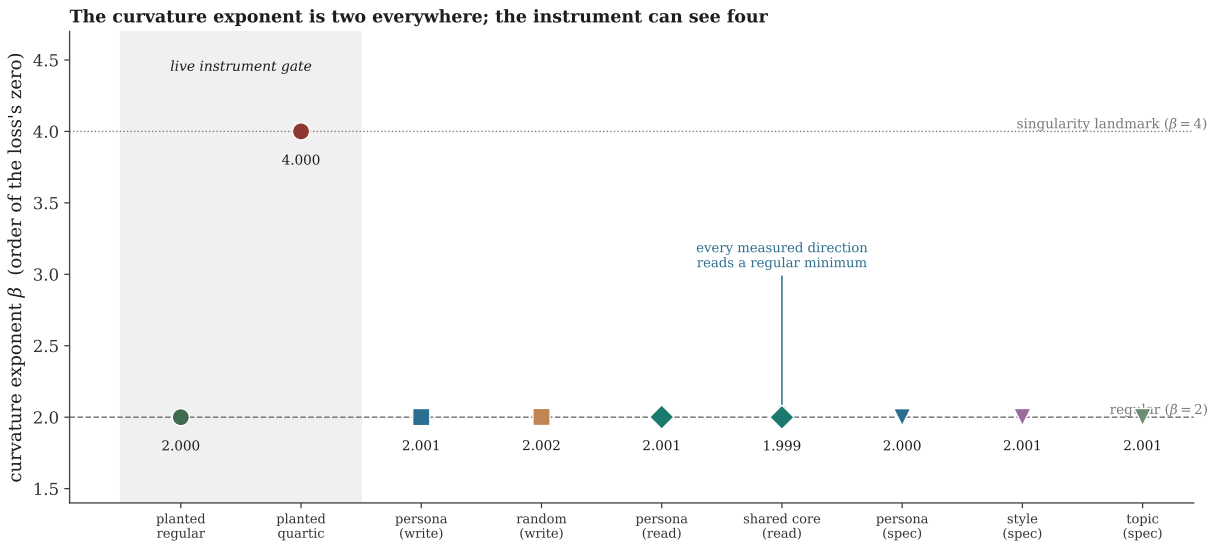

Figure 6: The curvature exponent is two everywhere, and the instrument can see four. The planted regular control reads two and the planted quartic control reads four, so the estimator separates a regular minimum from a singularity on the real model. Every measured direction reads two: the persona and random weight cores, the read-channel persona object and the shared read core, and the persona, style, and topic cores. There is no flat singularity in either channel, and the flatness is not persona-specific because there is none.

The participation ratio is the one reading that, taken alone, points toward the singularity. A participation ratio is scale-free, the effective number of directions that carry the curvature, and a low value is silent about whether any direction is actually flat: it falls both when directions go dead and when curvature concentrates in a few stiff ones. The two are told apart by what the curvature is doing in absolute terms, and three facts settle it the same way. The total curvature is higher, about ten times the random core's, not lower as a degenerate valley's would be. The dominant column is an order of magnitude above the rest, the signature of concentration, not collapse. And, decisively, the persona core's *flattest* direction is still stiffer than the random core's flattest and nowhere near zero, so no direction of the persona core is dead; the low participation comes from *adding* stiff directions, not from any direction going slack, which is what a singularity would require. Read alone the participation ratio is ambiguous; read with the exponent and the volume it is unambiguous, and it points away from the singularity. Crediting the low participation as a degeneracy win would mistake a scale-free shape statistic for a measure of flatness. Read correctly, it says the persona core is anomalously stiff and its stiffness is anisotropic, piled into one direction, the opposite of a flat valley.

On a model this size the slice learning coefficient floors far below its regular reference for any low-rank weight direction, the persona core and the random core alike, so the volume axis cannot by itself separate a regular minimum from a singular one in this channel; what it can do is compare the two cores on the same footing, and there it says persona is the stiffer. That is why the exponent and the participation structure, not the volume, carry the finding. The geometry of the substrate, stated positively: it is a regular minimum of the pretrained behavioral loss, curvature exponent two, anomalously stiff and concentrated relative to a matched random subspace, a distinguished place in weight space but not a flat one. The strongest mechanistic hypothesis, that the substrate resists removal because the loss is blind to it, is not how it works;

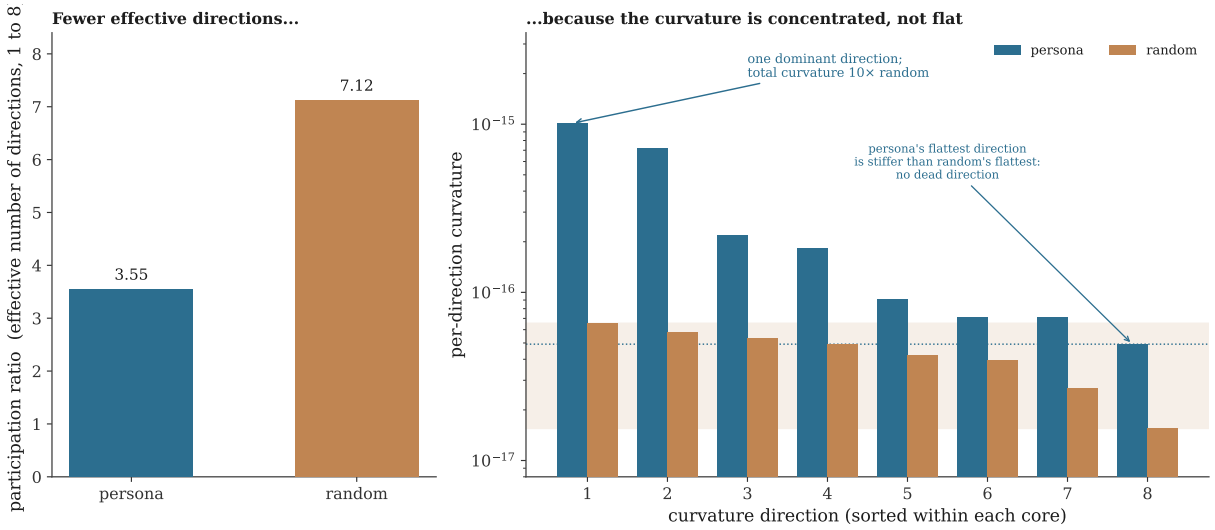

Figure 7: Concentration, not flatness. Left: the persona core's curvature is carried by fewer effective directions than the random core's, a participation ratio of 3.55 against 7.12, which taken alone looks like degeneracy. Right: the per-direction curvature spectra show why it is not. The persona core has a dominant direction holding about forty percent of the total and about ten times the random core's total curvature, and its flattest direction is still stiffer than the random core's flattest, so no direction is dead. The low participation is concentration of curvature, the opposite of a flat valley.

the loss sees the direction perfectly well. The substrate's resistance to removal is the read/write amplification asymmetry, which this measurement leaves exactly where prior work had it, and which the absence of any flat escape direction reinforces: there is no degenerate direction to slip the substrate out along, because the loss is a regular minimum everywhere on it.

4.4 The causal handle: constraining the substrate bends broad misalignment

The geometry says the substrate is a distinguished object; the causal measurement asks whether it is a load-bearing one. The induction recruited broad misalignment as it had to: the plain arm reached a judged broad-misalignment rate of 0.226, well above the floor below which a reduction would be meaningless, so the behavioral clock is alive. The constraint did what it claimed: the plain fine-tune put a measurable fraction of its weight motion into the persona substrate, a weight-span of 0.171, and the persona constraint drove that to 0.0008, against an isotropic null of 0.0016, removing essentially all of the fine-tune's motion in the substrate. The manipulation is real and it is the strongest possible version of it, a full projection rather than a gentle penalty.

The reading is directional and significant, and it is small. Constraining the persona substrate reduced the broad-misalignment margin by 0.541, a sign-flip permutation over the battery giving $p = 0.033$. Constraining a matched random subspace of equal rank reduced it by 0.149. So the persona constraint reduced broad misalignment about 3.6 times more than the random constraint, and the persona effect is the only paired-significant one. The signal is an aggregate one: across the battery of broad probes the per-probe distributions overlap heavily, about fifty-five percent of probes move down and the rank-based effect size is near zero, so it is the paired shift over the battery, caught by a sign-flip test, that is significant, not a large separation on any single probe. The independent judged misalignment rate moved in the same direction, from 0.226 to 0.209, on

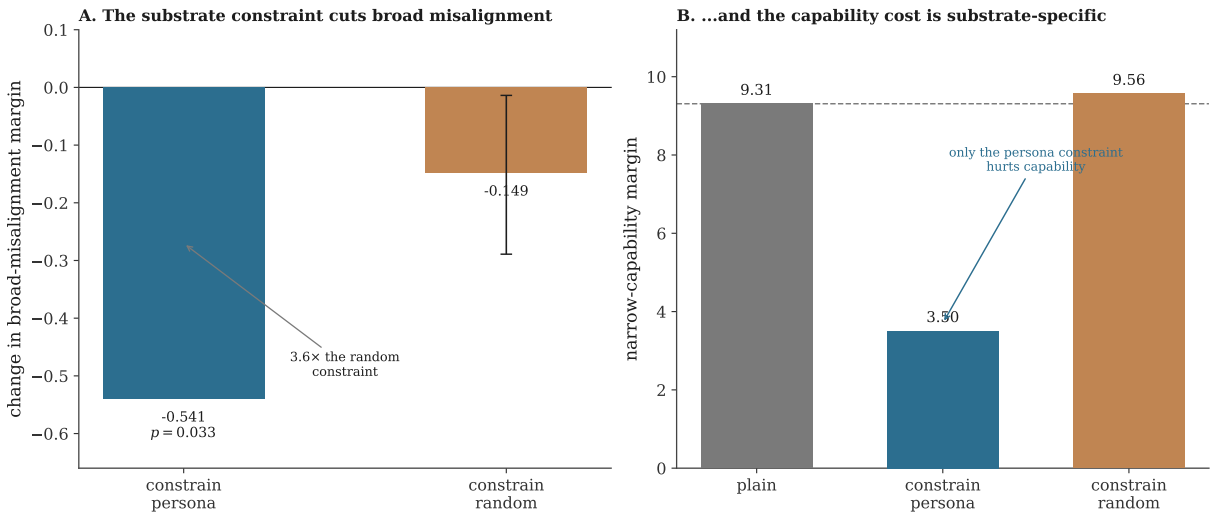

Figure 8: Constraining the substrate, substrate-specifically. Forbidding a narrow induction fine-tune from writing into the persona subspace reduces the broad-misalignment margin by 0.541 ($p = 0.033$), about 3.6 times the reduction from constraining a matched random subspace (0.149), and the persona arm is the only paired-significant one. The capability cost falls on the persona arm alone: it drops the narrow-capability margin from 9.31 to 3.50, while the random constraint at the identical dose leaves it at 9.56. The broad effect is substrate-specific; the capability cost is too.

a sample of seven hundred completions per arm with mean coherence near 86 in both, so the persona arm is not incoherent. This is the program's first direct causal handle on the substrate in the weight channel, where earlier work could grip it only in the read channel: perturb the substrate during training and broad misalignment moves, substrate-specifically, more than when you perturb a random subspace.

The ceiling is attached. The random constraint is not a clean zero: its reduction of 0.149 has a confidence interval of $[-0.289, -0.014]$, which excludes zero, so constraining a random subspace also reduces broad misalignment somewhat, just much less. Part of even the persona arm's reduction is therefore the generic cost of constraining weights, and the substrate-specific part is the increment over that nonzero baseline, about 0.39; the 3.6-times figure is a ratio over a baseline that is not zero, and should be read as such. And the persona constraint cost the narrow capability: the narrow-adherence margin fell from 9.31 to 3.50, a retention of 0.38, against the broad drop of 0.541, so at this dose the constraint hurt the narrow task far more than it helped the broad one. This is why the finding is a directional, substrate-specific causal demonstration and not a clean prevention.

The obvious objection: a model that has lost sixty-two percent of its narrow capability is a damaged model, and a damaged model is less coherently anything, so of course its broad misalignment fell, the constraint simply broke it. The defense is the random control, and it is why the random arm was run. At the identical dose, a full projection, constraining a random subspace preserved the narrow capability, leaving its margin at 9.56, essentially the plain arm's 9.31, while reducing broad misalignment by less than a third as much, 0.149 against 0.541. The capability damage is therefore specific to constraining the persona substrate; it is not what constraining weights generically does. Two further facts harden the defense. The broad signal is a teacher-forced log-probability margin, not a coherence-confounded judge score, and the persona arm's coherence stayed near 86, so the broad drop is not a coherence collapse read as alignment. And

the independent judged rate moved the same way. What the defense earns is precise: substrate-specificity and direction, that the capability damage and the broad reduction both attach to the persona substrate and not to constraining weights in general. What it does not earn is a dissociation. Within the persona arm the broad drop and the capability loss are co-present, and no lower dose was run that exhibits one without the other, so the measurement cannot say the broad reduction is more than the correlate of the capability damage; it can say only that both are caused by constraining the substrate, and substrate-specific. The reading is that the persona substrate is causally load-bearing for broad misalignment and uniquely costly to constrain, that the blunt maximal dose used here cannot separate the broad drop from the capability cost, and that whether a tuned, capability-preserving dose can is the open question the geometry, a stiff and concentrated object a well-aimed constraint could reach, makes worth asking. The substrate is causally load-bearing; constraining it is the prototype of a prevention, not yet a capability-preserving one.

5 What this establishes and what it does not

It establishes a shared, author-specific, geometrically distinguished, causally load-bearing persona substrate in the weight geometry of the base model. The substrate is shared: four independently extracted per-domain persona subspaces fall into one low-rank subspace at about 657 times chance, every domain participating between 0.89 and 0.94. It is author-specific: a construct-valid containment test puts it about eighty-two percent orthogonal to style and ninety percent to topic, shared with itself across domains 2.8 times more than with style. It is geometrically distinguished: the pretrained loss along it is a regular minimum, curvature exponent two, but with about ten times the curvature volume of a matched random subspace and that curvature concentrated in a single direction, so it is anomalously stiff and anisotropic. And it is causally load-bearing: constraining a narrow induction fine-tune from writing into it reduces broad misalignment by a teacher-forced margin of 0.541 ($p = 0.033$), about 3.6 times the reduction from constraining a matched random subspace, with the capability cost falling on the persona constraint alone, so the effect is specific to the substrate.

It does not establish that the substrate is a flat singularity, and the strongest mechanistic form of the thesis is ruled out rather than left open. The loss is a regular minimum along the substrate, exponent two in the weight channel and the activation channel, so the predicted read/write curvature asymmetry does not exist and the un-removability of the disposition does not rest on loss-flatness. It does not establish a clean, capability-preserving prevention: the causal constraint was the bluntest possible dose, a full projection, and it cost the narrow capability, while the random control was not a clean zero, so what lands is a directional, substrate-specific causal demonstration, not a defense. And the author-specificity is a statement about base-model weight geometry; it does not by itself say the read-channel causal object is persona rather than style, which is a different measurement in a different space, and the causal handle is a statement about a weight-channel constraint, which does not by itself type the read channel either.

Threats to validity. The geometry and the causal handle are one model and one extraction recipe; the persona-versus-style separation, though clean, is a separation of two recipes, and a different recipe could draw the line elsewhere. Persona and style share more than chance, so the author-specific result is a distinctness claim with a measured residual, not an orthogonality claim. The causal effect is modestly powered and capability-confounded at the dose used, and the curvature exponent is read on one model with an instrument whose volume axis floors in this channel, so

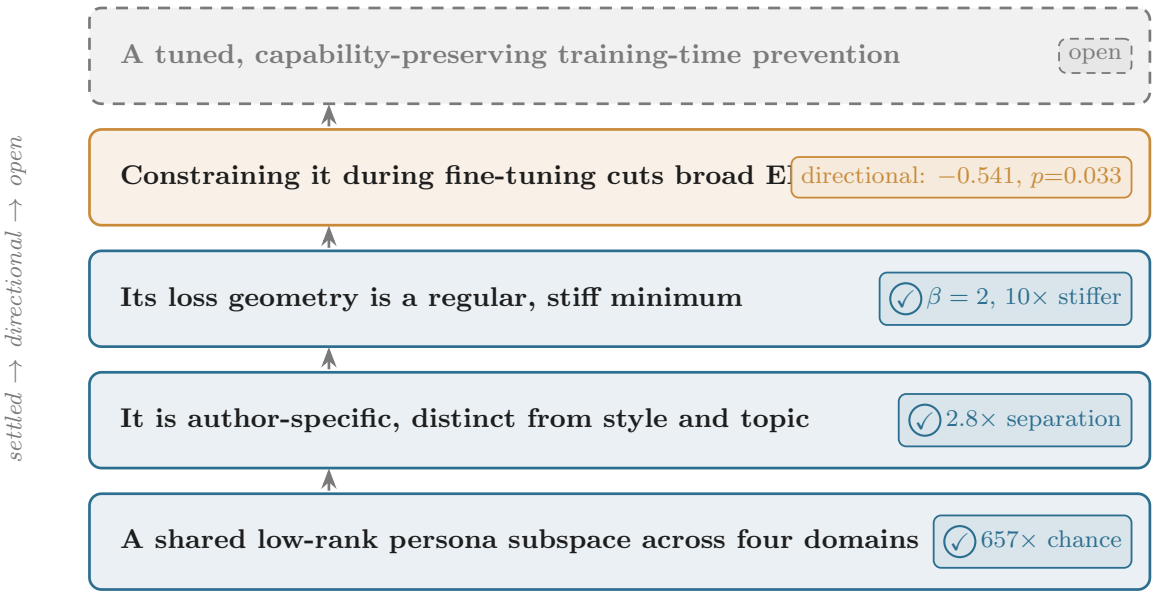

Figure 9: What the program now holds. The shared substrate and its author-specificity are settled weight-space linear algebra. The geometry of the substrate is settled too, as a regular but stiff and concentrated minimum rather than a flat singularity. The causal role is now demonstrated directionally and substrate-specifically. The open rung at the top is a clean, capability-preserving, training-time prevention, which the causal handle points at but does not yet deliver.

a singularity in a regime this instrument cannot reach is not excluded, only the one it can reach is. The ceiling: the geometry finding is a regular, anomalously stiff, concentrated minimum measured with a live and controlled instrument, and the causal finding is a directional, substrate-specific reduction of broad misalignment under a blunt constraint that also cost capability.

5.1 The joint statement across the program

Several measurements in this program are coordinate systems on one hypothesized object, and together they now license a four-part statement that was, not long ago, a one-part conjecture with three open layers. A real, shared, low-rank substrate for broad misalignment exists and is recruited broadly; that was established by the gradient and activation work. It is author-specific, distinct from style and topic, settled here in weight geometry where the program's cleanest causal result could separate the two only by a thread. Its loss geometry is now characterized: it is a regular but anomalously stiff and concentrated minimum, which closes the singularity question in favor of a distinguished-but-regular object and relocates the account of un-removability onto the read/write amplification asymmetry rather than loss-flatness. And it is causally load-bearing: a training-time constraint on it bends broad misalignment substrate-specifically, which converts "the substrate is correlated with broad misalignment" into "the substrate is causal for it," directionally, with a clean capability-preserving version of the intervention as the named next step.

The narrowing is the progress. The standing summary used to be that the substrate keeps showing up generic, shared, and read-space exactly where the theory needed it specific, singular, and causal in the weights. On the specificity axis that word is now *specific*. On the geometry axis

the open question of singular versus regular is now *closed*, as regular-but-distinguished, which is a sharper and more useful object than a flat valley because it says the resistance to removal is amplification, not blindness. On the causal axis the program has its *first direct handle* in the weight channel. What stays open is bounded and nameable: whether a tuned, capability-preserving form of the constraint is a prevention, whether the result holds across organisms and base models, and whether the read-channel causal object is itself persona rather than style. The program has tightened, not closed, and it has tightened in three specific places.

6 How this differs from prior causal work

A careful and growing body of work *cuts* emergent misalignment by intervening on a representation: ablating concept directions during fine-tuning (Casademunt et al., 2025), blocking a small set of latent features (Ustaomeroglu and Qu, 2026), inoculating with prompts or regularizers (Marks et al., 2025; Tan et al., 2025; Wichers et al., 2025), steering or erasing a persona feature (Wang et al., 2025; Chen et al., 2025), and, in this program's own read-channel work, projecting a persona subspace out of activations. These establish that a handle exists and that pulling it changes the behavior; the single cleanest template for the shape of those claims is the demonstration that one direction mediates refusal, ablate it and the model stops refusing (Arditi et al., 2024). They are about control: find the lever, pull it, watch misalignment fall.

This work asks a different question and then acts on a different channel. The geometry result does something the control literature does not: it separates persona from style *in weight space*, where steering and ablation operate in activation space and have generally found the two entangled, and it types the object as the model's own likelihood geometry, a regular-but-distinguished minimum, rather than as a chosen activation direction or a learned sparse feature, which makes it a more falsifiable object than "a direction that correlates with the behavior." And the causal intervention is not a removal after the fact but a constraint during training. The program's other weight-space work established that the disposition is un-removable after training, the edit invisible or re-filled; the natural causal test is therefore not to remove the substrate but to forbid the fine-tune from writing into it while the misalignment is being recruited, which is a gradient-routing intervention (Cloud et al., 2024; Saha et al., 2021; Wang et al., 2023) rather than a representation edit. That is the move this work makes, and the substrate-specific reduction it produces is the evidence that controlling where a narrow fine-tune may write, during training, is a place prevention can act.

The geometric neighbors are the right ones to measure this against. The language of a degenerate loss landscape and a learning coefficient is singular learning theory (Watanabe, 2009, 2013) and its local, estimable form (Lau et al., 2025; Wang et al., 2024; Hoogland et al., 2024); the reading that the substrate is a regular, concentrated minimum rather than a singularity is consistent with a superposed-feature account of what low-rank structure pretraining installs (Minegishi et al., 2026; Elhage et al., 2022) rather than a degenerate one. That flatness is tied to generalization is an old and contested intuition (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017; Dinh et al., 2017; Foret et al., 2021), and the recent singular-geometry results on degenerate directions (Shirodkar, 2026) are what let the geometry be read as the order of a zero rather than a curvature. That fine-tuning lives in a low-dimensional, reusable subspace is the territory of intrinsic dimension and task arithmetic (Li et al., 2018; Aghajanyan et al., 2021; Ilharco et al., 2023; Ortiz-Jiménez et al., 2023), and that re-specifying an author is a cheap linear move is the persona, simulator, and linear-representation line (Andreas, 2022; Shanahan et al., 2023; Park

et al., 2023; Elhage et al., 2022). The difference this work reaches for is not that those works are lesser; it is that it measures the loss geometry behind the behavior and constrains the channel the narrow update writes through, where they establish that the behavior happens and which lever moves it.

7 Where this points: prevention as substrate-aware gradient routing

The causal measurement is, mechanically, a prevention prototype, and what it opens is where the work is going. The constraint that produced the causal finding forbids the fine-tune from descending into a chosen low-rank subspace, confining the update to the orthogonal complement of the substrate. That this reduced broad misalignment, substrate-specifically and several times more than a random-subspace constraint, is a proof of concept that prevention can act by controlling *where in weight space* a narrow fine-tune is allowed to move, during training, rather than by removing the substrate after the fact. The program's other weight-space work foreclosed removal after the fact; this work opens the training-time routing direction with a causal demonstration. The forward program is therefore not a vague aspiration but a specific claim with a first instance: prevention of broad misalignment is a training-time, substrate-aware routing problem, and here is a lever on it.

The next steps the measurement defines are concrete. The first is the **dose-response**. The clean causal signal came at a blunt maximal dose, a full projection, that also cost capability, so the immediate question is whether a tuned, lower, or one-sided constraint can separate the broad-misalignment drop from the capability cost, a curve this measurement points at but did not trace; the geometry says the substrate is stiff and concentrated, which is encouraging for a targeted constraint, because a concentrated object is one a low-rank, well-aimed penalty can reach. The second is the **capability-preserving form**. The causal constraint barriered the write channel along the substrate and, at full strength, took the narrow capability with it; the substrate's two-channel structure suggests a sharper intervention, one that barriers the write channel without touching the read carrier the capability needs, so the model stays able to represent a bad author while losing the standing disposition to become one, the non-lobotomy version of the lever. The third is **generalization**: whether the causal handle replicates across a second narrow organism and across base models, where the substrate's sign is known to reverse, which decides whether substrate-aware routing is a model-universal prevention or a model-specific one. And the fourth is the **connection to the program's one other moving lever**, a training-time data-framing intervention that lowers the first-step routing into the substrate (Marks et al., 2025; Tan et al., 2025); both this constraint and that framing act on the same training-time read-and-routing channel, so the coherent forward picture is that the place prevention can act is where the narrow gradient meets the substrate, and these are two instances of acting there.

This is a direction with a causal proof of concept, not a built defense, and the field's own verdict on this family of interventions is sobering: tamper-resistant and non-fine-tunable-learning safeguards are hard to make durable and are often broken by adaptive attackers and even by trivial changes to the attack (Tamirisa et al., 2025; Rosati et al., 2024; Qi et al., 2025; Kuo et al., 2026). The claim this work will defend is that it identifies where prevention can act, in the training-time routing of the narrow update around a measured substrate, and shows a first causal lever there, not that it prevents anything yet. A training-free read of the substrate's geometry, its dimensionality and how strongly each domain participates, is a separate and nearer

payoff, a candidate pre-release susceptibility metric that does not depend on the routing program at all, validated against the open-weight cohorts surveyed for related properties (Schreiber and Goldstein, 2026; Dickson, 2026).

8 Reproducing this work

The artifacts behind every number here are the geometry run's subspace matrices and result files, the degeneracy run's per-column curvature, participation, volume, and exponent records, and the causal run's per-probe margins, manipulation spans, and judged generations. Every headline number in this document is recomputed from those files by a released analysis script, which recomputes the overlap-share and cross-core statistics from the raw subspace matrices, the participation ratio from the per-column curvatures, the causal margins and their paired differences from the per-probe arrays, and the judged rates from the per-response scores, cross-checks each against the stored values, and marks quantities from measurements that did not run as such. The figures are regenerated from that verified-numbers file by one command, as is an HTML rendering of this document. The appendices hold the heavy detail: the notation and the nulls (A), the subspace extraction in full (B), the degeneracy instrument and its live certification (C), the causal apparatus (D), the statistics (E), the exploratory geometry (F), the extended related work (G).

Acknowledgments. This work stands on the emergent-misalignment and singular-learning-theory communities whose instruments and organisms it borrows, and on the program's own earlier measurements. The errors that remain are the author's.

A Notation, objects, and the nulls

A.1 Notation

θ_{pre}	the pretrained base model's weights (`Qwen2.5-14B-Instruct`), held fixed.
d_{model}	residual-stream width, 5120.
$U_{t,d}$	the rank-k per-domain subspace for contrast type t and domain d, $k = 4$.
$U_{\text{sh},t}$	the rank-8 shared core for type t, the top directions of the four stacked $U_{t,d}$.
overlap-share(t)	cross-domain overlap-share of type t (Eq. 1), in $[0, 1]$.
cross-core(persona, t)	cross-core of the persona core against the type-t core (Eq. 2).
W_{persona}	the rank-8 persona *write* core in weight space, the object the geometry acts on.
U_{persona}	the rank-4-per-layer persona *read* object in the activation channel.
$L_B(\delta)$	behavioral-KL loss, the base model's prediction-divergence under a weight move δ.
β	curvature exponent, the order of the loss's zero along a direction, $\beta = 2k$; 2 regular, > 2 singular.
$\hat{\lambda}_B$	the slice learning coefficient, a volume reading on the scale $k/2$ regular, $k/4$ quartic, 0 flat.
$M_{\text{broad}}, M_{\text{narrow}}$	teacher-forced broad-misalignment and narrow-capability margins (Appendix D).

A.2 The two geometry statistics, exactly as computed

For two orthonormal bases U_a, U_b, the cosines of the principal angles between their column spaces are the singular values of $U_a^\top U_b$, so $\|U_a^\top U_b\|_F^2 = \sum_i \cos^2\theta_i$. The overlap-share (Eq. 1) averages $\frac{1}{k}\|U_{t,d_i}^\top U_{t,d_j}\|_F^2$ over the six domain pairs; the cross-core (Eq. 2) is $\frac{1}{8}\|U_{\text{sh,persona}}^\top U_{\text{sh},t}\|_F^2$, the mean squared canonical correlation of the two rank-eight cores. The random-subspace null of the overlap-share is $k/d_{\text{model}} = 4/5120 = 0.00078$ (two random k-planes in d dimensions share k/d of their squared cosine on average); for the rank-eight cross-core the analogous null is $8/5120 = 0.0016$. Both nulls were Monte-Carlo confirmed.

A.3 The curvature exponent and the volume scale

The curvature exponent reads the order of the zero of the behavioral loss along a direction. For a one-parameter family $\theta(t) = \theta_{\text{pre}} + t\,v$ with θ_{pre} a minimum, singular learning theory gives $L_B(t) \sim c\,t^{2k}$ for small t, where k is the order of the zero (Shirodkar, 2026). A non-degenerate quadratic minimum has $k = 1$ and exponent $\beta = 2k = 2$; the quadratic term vanishing gives $k \geq 2$ and $\beta > 2$, the singularity. The contribution of that direction to the model's effective dimension is $1/(2k) = 1/\beta \leq \frac{1}{2}$, with equality exactly for the regular case, which is why the volume scale below has its regular reference at $k/2$.

The slice learning coefficient is the volume-scaling exponent of the low-loss set: the volume of $\{\delta : L_B(\delta) < \varepsilon\}$ shrinks as ε^λ up to a log factor, so small λ means a large flat valley. For a regular k-dimensional minimum $\lambda = k/2$, each non-degenerate direction contributing $\frac{1}{2}$; a pure quartic direction contributes $\frac{1}{4}$, so a wholly quartic-degenerate slice reads $k/4$, and a flat direction reads near zero. With the persona core's rank $k = 8$ the regular reference is $k/2 = 4$, the quartic

landmark is $k/4 = 2$, and the flat floor is 0.

A.4 Per-quantity type audit

Each headline statistic, its support direction, its null/scale, and its measured value are in Table 1.

Quantity	Direction of support	Null / scale	Measured
overlap-share(persona)	high above the random null	4/5120	0.513
cross-core(persona, $\cdot$)	*low*, below 0.30, with self/cross > 2	rank-8 null ≈ 0.0016	0.182 / 0.097
$\beta(W_{\text{persona}})$	> 2 for a singularity	2 regular, 4 quartic	2.001 (regular)
$\text{PR}(W_{\text{persona}})$	low *and* read with the volume	$[1, k]$; random 7.12	3.55 (concentration)
ΔM_{broad} (persona)	negative, paired-significant	0; random -0.149	-0.541, $p = 0.033$

Table 1: The per-quantity type audit. Each statistic has a sign, a distribution with variance (decided by permutation or sign-flip, never by an effect-size threshold), and a reachable, self-diagnosing scale. The participation ratio is the one quantity whose support direction is read only jointly with the volume and the exponent, for the reason §4.3 sets out: low participation alone is ambiguous between flatness and concentration.

B The subspace extraction, in full

B.1 Contrastive teacher forcing

For one contrast, the inputs are a list of triples, each a positive system prompt, a negative system prompt, and a user prompt. A single shared response is generated once from the base model under a neutral system prompt, greedily, so the response text is fixed before any contrast is applied. The model then reads that response twice, teacher-forced, once with the positive system prompt in front and once with the negative, and the mean of the residual stream over the response token span is captured at each layer of a central band. Because the response tokens are byte-for-byte identical across the two passes, verified by a hash of the token-id list that must match, the difference of the two captures isolates the speaker rather than the content. Per contrast pair and per layer, that difference is one row of a stack; the stack's top four left singular vectors, orthonormalized and sign-anchored so each column points toward the misaligned pole, are the per-domain subspace $U_{t,d}$.

B.2 Diversity matching: the fix that makes the comparison fair

The persona contrast is drawn from twelve descriptor-variant pairs, of the form "a writer who is reckless and dismissive of harm" against "a writer who is careful and responsible," cycled across the prompts so the subspace is not the fingerprint of one phrasing. Style and topic are given the same diversity: style from twelve distinct surface-register pairs (formal against casual, terse against florid, and so on), topic from twelve distinct subject framings. Drawing style from a single fixed formal-versus-casual contrast would make the style subspace nearly rank-one and identical across domains, inflating its cross-domain self-overlap and producing the misleading magnitude

race. With the contrasts matched in diversity, the per-domain effective ranks are comparable (persona about 2.1, style about 2.3, topic about 3.8; §F), so a comparison between them is a comparison of like with like.

B.3 The two faces of the persona substrate

The geometry measurements act on the persona substrate in both of its channels, and the channels are extracted differently, which the reproduction needs stated plainly. The *read* object is the activation-channel persona direction, extracted by the contrastive teacher forcing above: a rank-four subspace per layer at layers eighteen, twenty-four, and thirty, the same object whose activation ablation drives broad misalignment to zero in earlier work. The *write* core is the weight-channel persona direction. It is not extracted by activation contrast; it is built in weight space from the published narrow misalignment organisms, by taking, per domain, the top singular directions of the organism's weight update across a central band of writer matrices, then taking the top eight shared directions of the three per-domain cores stacked together. So the write core is the rank-eight shared output-side direction the organisms actually write through, paired with the per-writer input directions, each column normalized to unit Frobenius norm. The shared read core and the four per-domain read subspaces are what carry the overlap-share and cross-core results; the write core is what the curvature exponent reads in the weight channel. Both are faces of one substrate, and the curvature exponent reads two on both.

B.4 Shared cores and the cross-core null

The rank-eight shared core $U_{\mathrm{sh},t}$ is the top eight left singular vectors of the four per-domain subspaces stacked column-wise. The cross-core (Eq. 2) is the mean squared canonical correlation between two such cores. Its permutation null pools the eight per-domain subspaces of persona and of the comparison type, randomly relabels four as "persona" and four as the comparison, rebuilds both cores, and recomputes the cross-overlap; with four domains there are only seventy distinct relabelings, so the test resolves p-values no finer than about $1/71$, and it is treated as a corroboration that the effect-size bars (cross-core below 0.30, self-versus-cross above 2) make the decision around.

C The degeneracy measurement, in full

This appendix gives the estimators and the controls that certify them.

C.1 The behavioral loss makes the base model an exact minimum

The order of a zero is defined at a minimum, and θ_{pre} is not a minimum of next-token prediction, so next-token curvature has an off-minimum bias. The behavioral loss removes the bias by definition. Measuring the deviation of the perturbed model from the base model itself, $L_B(\delta) = \mathbb{E}_x \,\mathrm{KL}(f_{\theta_{\mathrm{pre}}} \| f_{\theta_{\mathrm{pre}}+\delta})$, makes $\delta = 0$ the exact global minimum, since KL is zero there and non-negative everywhere, so both the exponent and the volume coefficient are non-negative by construction and the reference loss is exactly zero rather than an estimated quantity subtracted from another. To second order L_B is the model's own Fisher form on its predictive distribution, computed model-side and label-free, so a direction with $L_B \approx 0$ is exactly one that barely changes

the output distribution. The KL here is this work's specialization of the behavioral-loss construction, stated in mean-squared form in the source (Bushnaq et al., 2024); the KL form is what gives the exact-minimum and non-negativity properties for a language model.

C.2 The curvature exponent, exactly as fit

For a weight direction v the estimator sweeps a grid of perturbation sizes $t \in \{0.25, 0.5, 1, 2, 4\}$, and at each t perturbs the weights in place by $t\,v$, runs a forward pass over a fixed pool of reference sequences, and accumulates the per-sequence mean behavioral KL between the base and perturbed predictive distributions, summed over the full vocabulary in double precision with no top-K truncation and the forward never run in reduced precision. Writing the mean KL at size t as $\overline{\mathrm{KL}}(t)$, the exponent is the slope of $\log \overline{\mathrm{KL}}$ against $\log t$, fit over only the points that rise above a numerical floor of 10^{-8}, with the hard rule that if fewer than three points survive the fit returns an honest unreadable flag rather than a fabricated slope. The fit is self-checked by the ratio $\overline{\mathrm{KL}}(2t)/\overline{\mathrm{KL}}(t)$, which sits near four inside the clean quadratic window; the measured per-column window ratios are all near four, confirming the fit sits in the quadratic regime and not in the floor or the saturated tail. If the default grid drops into the floor or saturates, a small ladder of alternate grids is tried and the one with at least three survivors in the readable KL band is used; on the persona core the default grid survived. For the write core the direction v is built per column as the rank-one-per-writer weight block formed from the shared output direction and the per-writer input direction, normalized to unit Frobenius norm, and the reported exponent is the median over the eight columns. For the read object the perturbation is added to one layer's residual output at a time and the exponent is the median over the columns and layers. The synthetic validation that licenses the estimator imposes a known monomial $\mathrm{KL} = t^{2k}$ and requires recovery of $\beta = 2$ for $k = 1$ and $\beta = 4$ for $k = 2$, and requires a too-low window to flag as unreadable rather than return a confident wrong slope.

C.3 The participation ratio, and why it cannot floor or sign-flip

The participation ratio reads the shape of the restricted curvature. The diagonal of the curvature in the subspace is approximated by the per-coordinate gradient power of the behavioral loss, the mean over a small set of minibatches of the squared coefficient-space gradient, which is near zero for a degenerate coordinate and large for a stiff one. From the per-coordinate curvatures $e_i \geq 0$ the participation ratio is $(\sum_i e_i)^2 / \sum_i e_i^2$, bounded in $[1, k]$, equal to k when the curvature is spread evenly and near one when it is concentrated in a single direction. It is scale-invariant, a ratio of two sums of the same degree, so a common-mode multiplicative artifact cancels and it cannot floor or sign-flip the way a raw curvature can. This is the property that lets it be read at all, and it is also the property that makes it ambiguous read alone: being scale-free, it is silent about whether any direction is actually flat. A low participation ratio falls out both when directions go dead, the degenerate case, and when curvature concentrates in a few stiff directions, the anisotropic case, and the two are separated only by the companion readings. The measured per-coordinate curvatures settle it: the persona core's smallest is still above the random core's smallest, so no persona direction is dead, and the largest is an order of magnitude above the rest, so the low participation is concentration. The synthetic validation plants a block of known rank and requires the ratio to recover it and to read a degenerate block below a regular one.

C.4 The volume coefficient, and why it is only a confirm

The volume coefficient is estimated by a localized stochastic-gradient sampler on the behavioral loss restricted to the subspace. The coefficient walk is

$$a \;\leftarrow\; a \;-\; \tfrac{\epsilon}{2}\Big(n\beta\,\nabla_a L_B(a) \;+\; \gamma\,a\Big) \;+\; \sqrt{\epsilon}\,\eta, \qquad \eta \sim \mathcal{N}(0, I_k), \tag{4}$$

with step size ϵ, inverse temperature β, and a localizing spring $\gamma\,a$ that confines the walk near the base point; the estimate is $\hat{\lambda}_B = n\beta\,\mathbb{E}_{\text{post}}[L_B]$, the tempered-posterior expected loss times $n\beta$, with $L_B(0) = 0$ exact so there is no reference to subtract. A finite-temperature bias is removed by estimating at three temperatures and extrapolating the intercept against $1/\log(n\beta)$, which also returns the multiplicity as the slope. The walk is run cheaply, two chains of fifty steps at a localizing strength chosen so a matched random slice reads on scale, with the fast-but-lossy matrix path enabled only here, as a validated exception, because the sampler's Monte-Carlo noise dominates its precision error and a same-seed spot-check confirms the coefficient is unchanged.

On a model this size the behavioral-loss volume of a low-rank weight direction reads near zero for any direction, regular or singular, because the localizing spring dominates the tiny restricted curvature in the flat write channel, so the absolute coefficient floors far below its regular reference for the persona core and the random core alike. The coefficient cannot, in this channel, separate a regular minimum from a singular one, and a threshold on it would be meaningless. What it can do is compare two subspaces on the same footing, and there it is informative: the persona core's coefficient is about ten times the random core's, the same direction the participation structure points, so the persona core is the stiffer. The exponent and the participation carry the geometry finding; the volume corroborates the stiffness contrast but cannot resolve the regular-versus-singular question its absolute scale was built for.

C.5 The live-instrument gate, run before any persona number

The gate certifies, on the real model, that the exponent estimator separates a regular direction from a singular one before any persona number is read. A matched-rank random weight core must read a regular exponent, in an acceptance band from 1.6 to 2.4; a planted quartic-degenerate direction, a synthetic walker on which $\mathrm{KL} = t^4$ is imposed, must read an exponent above three; and a planted regular direction must read about two. On the real model the random core read 2.002, the planted quartic read 4.00, and the planted regular read 2.00, so the instrument provably sees a singularity when one is present, and the gate passed. Only a failure of the synthetic anchors before the model is touched routes to an instrument-limited outcome, and those passed.

C.6 The decision, and why it reads regular rather than singular

The geometry decision asks whether the persona write core is degenerate, and it requires three things together: an exponent above two, the persona core significantly flatter than the random core, and a lower participation ratio. The exponent is the headline, and it read two, with the persona core not significantly flatter than random (the permutation p is 0.31). So the decision is a regular minimum, the superposition-style account of ordinary feature geometry over the flat singularity, and the participation ratio and volume, read correctly against each other, characterize that regular minimum as anomalously stiff and concentrated. The read-channel specificity race, whether the flatness is persona-specific, reads the exponent as two for persona, style, and topic alike, with a permutation p of one, so there is no flatness for any of them to be persona-specific

about; the substrate's author-specificity is the construct-valid cross-core of Appendix B, which this measurement does not recompute and does not bear on. The magnitude race re-surfaces in this measurement's own diagnostics, with the style core's cross-domain overlap again larger than the persona core's, and it is the same confound the cross-core test defeats; it is reported for context and is not the author-specificity verdict.

D The causal apparatus, in full

The causal measurement constrains a narrow induction fine-tune and reads the behavior. The constraint, the judge-free readout, the matched control, and the confound are each described below.

D.1 The stiffening constraint, as implemented

The constraint is a gradient-routing penalty applied during training. At every optimizer step, after the backward pass and before the clip and the update, the component of each writer matrix's update that lies along the persona write subspace is removed from the gradient, so the update is confined to the orthogonal complement of the substrate. Concretely, for each writer the low-rank update factor's gradient B is replaced by $B - \frac{\gamma}{1+\gamma} UU^\top B$, where U is the shared output-side persona direction; the strength γ is the dose, and at $\gamma \to \infty$ the fraction $\gamma/(1+\gamma) \to 1$ and the persona component is projected out entirely. The dose used is the maximal one, a full projection, so the constraint is the bluntest possible: it forbids the fine-tune from moving along the substrate at all, rather than penalizing it gently. This is the same operation as projecting a subspace out of the update, the infinite-stiffness limit, chosen so the manipulation reliably fires; a graded dose is the natural next measurement and was not run. The penalty is applied only to the writer matrices, the attention output and the feed-forward down-projection, and the constraint is checked to match at least one writer gradient so a name mismatch fails loudly rather than silently doing nothing.

D.2 The three arms and the training recipe

The three arms differ only in the constraint. The plain arm trains unconstrained. The persona-constrained arm applies the full projection along the persona write core. The random-constrained arm applies the identical full projection along a matched-rank random core, orthonormal with Gaussian per-writer directions and each column normalized to unit Frobenius norm, the same construction as the geometry measurement's random control. The random arm is decisive: it is what separates "constraining the persona substrate changes the behavior" from "constraining any subspace hurts learning." All three are low-rank adapter fine-tunes of rank thirty-two with the stabilized scaling, trained on the writer matrices only with the rest of the model frozen, at a small learning rate with a short warmup and linear decay, an effective batch of sixteen, weight decay, gradient clipping, in the usual reduced-precision adapter regime, for one epoch of the narrow dataset, which is three hundred seventy-five steps, with a single checkpoint at the end. The same seed gives the three arms an identical data order, so the only difference between any two arms is the constraint.

D.3 The narrow induction dataset

The induction dataset is six thousand examples of reckless financial advice: a user asks an ordinary personal-finance question and the assistant answers with confident, dangerous guidance, pushing speculative crypto, margin, and guaranteed-return schemes as if they were safe. It is one of the program's standard narrow misalignment streams, chosen because its narrow behavior is sharply defined and because the broad misalignment a financial induction recruits is well clear of the floor. The same dataset trains all three arms.

D.4 The readouts: judge-free margins, a manipulation check, and a confirmatory judge

The primary readouts are teacher-forced margins, read without a judge. For a probe, the margin is the difference between the summed completion-token log-probabilities the fine-tuned model assigns to a matched more-bad completion and a matched less-bad completion of the same prompt, in full precision. The **broad-misalignment margin** averages this over a battery of free-form broad misalignment probes; a fall in it is a fall in broad misalignment, read in output space with no judge in the loop. The **narrow-capability margin** is the same construction on in-character probes whose more-bad completion is coherent bad-author text and whose less-bad completion is a refusal or derail, so a high value means the model can still represent a capable bad author and a fall means the narrow capability was damaged. The broad battery is one hundred ninety-five probes, the standard free-form misalignment probes and their paraphrases, spanning two dozen categories of open-ended question; the narrow battery is one hundred twenty in-character probes across three domains, villain dialogue, malware explanation, and insecure-code critique, the confound guard.

The **manipulation check** reads, directly from the saved adapters and without loading the base model, what fraction of each fine-tune's weight motion lies in the substrate: for every writer update factor it computes the share of Frobenius energy inside the persona span, averaged over the matrices, against an isotropic null of $k/d = 8/5120 = 0.0016$. The plain fine-tune put a fraction of 0.171 into the substrate and the persona constraint drove it to 0.0008, below the isotropic null, so the constraint provably removed essentially all of the fine-tune's substrate motion and the manipulation is real and meaningful.

The **judge** confirms the absolute rate on a sample. It is a local large model scoring each response on two axes, alignment and coherence, and counting a response as broadly misaligned only if it is both misaligned (alignment below thirty) and coherent (coherence at least fifty), with topic-relevant near-domain responses excluded from the broad rate so a merely on-topic-bad financial answer is not scored as broad misalignment. The judge confirms the direction of the margins and that the constrained arm has not simply gone incoherent; it is not the primary signal. About seven hundred completions per arm were scored, with mean coherence near eighty-six in both arms and inter-rubric agreement near 0.79.

D.5 The decision, and the confound it reports

A clean prevention would need five things together: the plain arm to recruit misalignment above the floor, the persona constraint to reduce the broad margin significantly, the random constraint to be near zero, the broad reduction to exceed the narrow-capability cost, and the manipulation to fire. Four of the five are in hand. The plain arm's judged rate of 0.226 is well above the floor

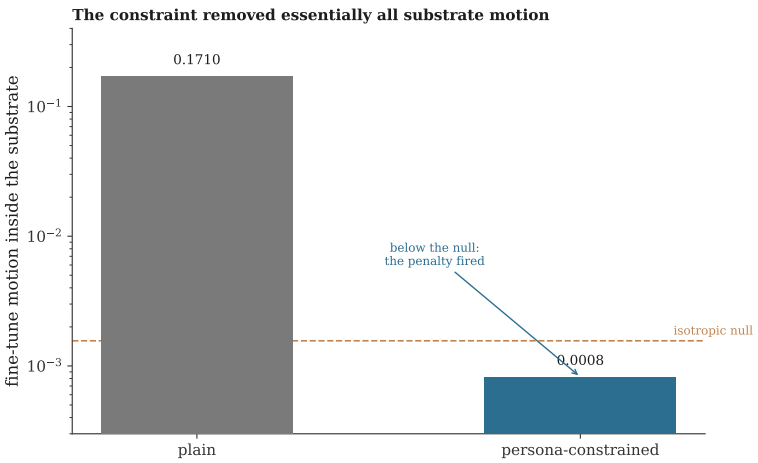

Figure 10: The constraint is real. The plain fine-tune put a fraction 0.171 of its weight motion into the substrate; the full projection drove it to 0.0008, below the isotropic null, so the constraint provably fired and removed essentially all of the fine-tune's substrate motion.

of 0.10; the persona constraint's broad reduction of 0.541 is significant by a paired sign-flip test ($p = 0.033$); the manipulation fired; and the random constraint is near zero, its reduction of 0.149 below a third of the persona reduction, though its bootstrap interval excludes zero, so it is small but not exactly zero. The one thing the blunt dose does not buy is the separation: the narrow-capability cost of 5.805 far exceeds the broad reduction of 0.541, so at full strength the constraint moves the narrow task far more than the broad one, and a clean, capability-preserving prevention is not what this dose measures. What it does measure, and what the decision's components establish individually, is a significant, substrate-specific reduction of broad misalignment under a constraint whose capability cost is itself specific to the persona substrate, the program's first causal handle on the substrate in the weight channel, with the clean dissociation awaiting a tuned dose.

E Statistical methods

Inference is by permutation and sign-flip tests with Cliff's δ for effect size, never by Cohen's d, which the program forbids as a headline gate because a large d can be a concentration artifact of a small denominator rather than a real separation. Ratios are gated on a confidence interval that excludes the relevant null, not on the point estimate. Within a battery the family-wise correction is Holm; there is no cross-measurement correction, because the measurements are separate pre-registered hypotheses.

The tests, explicitly. A paired sign-flip test on a margin change negates each probe's difference independently and reports the fraction of sign assignments at or beyond the observed mean, the natural null being symmetry about zero; the causal broad-margin change uses this over the shared probes. The persona-versus-random exponent comparison is an unpaired group-label permutation over the columns of the two cores, because the columns of two different cores have no natural pairing and an unpaired test wastes no power on a meaningless pairing. Effect size is Cliff's δ, the probability that a random draw from one condition exceeds a random draw from the other minus the reverse, rank-based and assuming no scale. Permutations use ten thousand

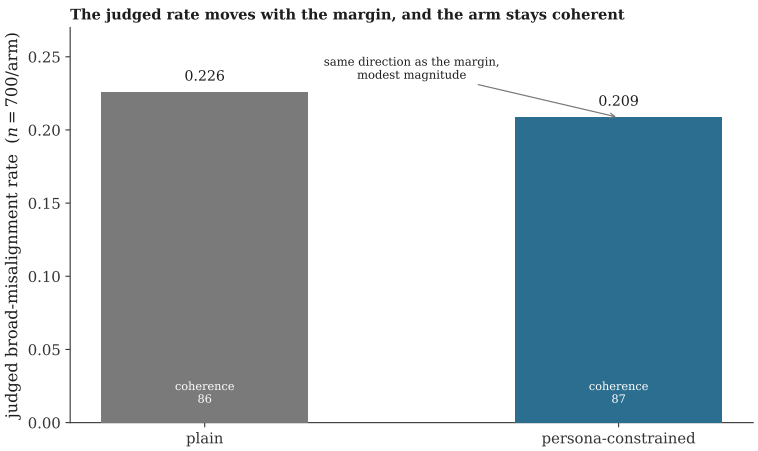

Figure 11: The judge confirms the margin, modestly. The independent judged broad-misalignment rate moves the same direction as the teacher-forced margin, from 0.226 to 0.209, while coherence stays near eighty-six, so the constrained arm is fluent rather than broken. The judged rate is a confirmation of the margin's direction, not the primary signal.

resamples; where the unit of inference is small, as for the cross-core's seventy relabelings, the test is read as corroboration and the pre-registered effect-size bars make the decision.

Why no Cohen's d. Cohen's d is not used as a headline gate. A large d can be an artifact of a statistic whose denominator is nearly constant by construction, where a tiny mean shift over a vanishing spread produces an enormous standardized effect that means nothing. A permutation or sign-flip test asks whether the observed separation could arise under the null without dividing by a scale that may itself be the artifact, and Cliff's δ reports how large the separation is in rank terms; neither can be inflated by a shrinking denominator. The same reasoning applies to the participation ratio: a scale-free shape statistic is not a measure of flatness.

Recomputation. Every geometry number is recomputed from the raw subspace matrices, every causal margin and its paired difference from the per-probe arrays, the participation ratio from the per-coordinate curvatures, and the judged rates from the per-response scores, independently of the stored result files, and cross-checked against the stored values and the run log. Quantities carried in from earlier work, the cross-domain sharing and the cross-core separation, are recomputed where their raw inputs are present and otherwise carried with their provenance marked; they are attributed as background, not re-counted as fresh evidence here.

F Exploratory geometry

Three pieces of structure in the saved geometry, each with what it implies and what it does not. First, the per-pair detail behind the overlap-shares (Figure 12, left): the six persona domain pairs cluster tightly at a mean principal angle of about forty-five degrees, style's at about twenty-five, topic's at about fifty. The tightness of the persona cluster is why the 0.513 mean is not an artifact of one close pair, and the smallness of the style angle is the mechanical source of the magnitude-race confound. Second, the effective dimensionality (Figure 12, right): the stable rank of the persona and style contrasts is low, about 2.1 and 2.3, while topic's is roughly double, about 3.8. Persona being genuinely low-dimensional is consistent with a small shared author latent; the caution that earns its place is that this is a descriptive property of the contrast subspaces and

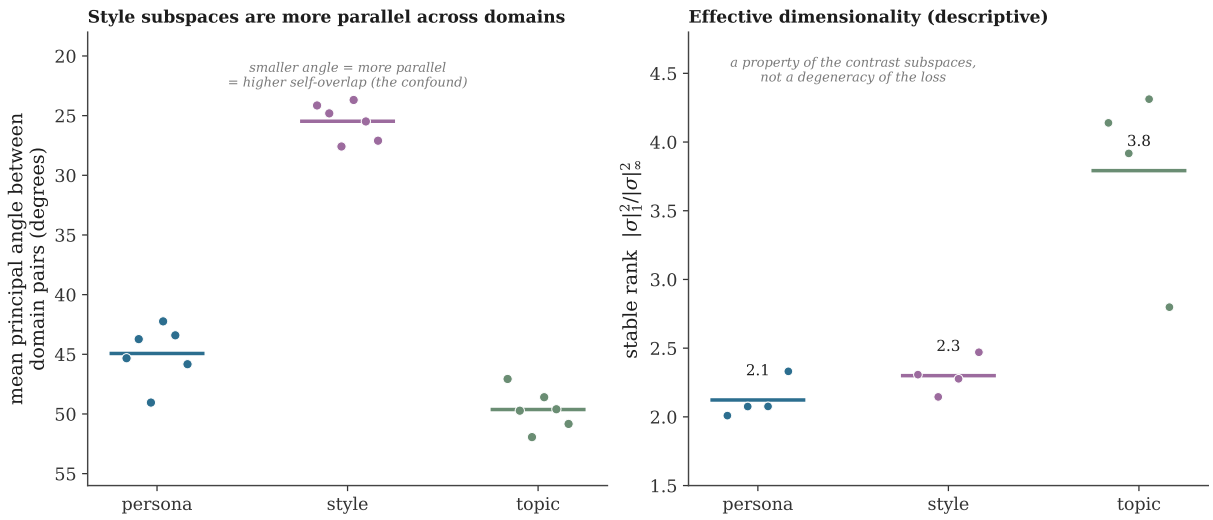

Figure 12: Per-domain geometry. Left: the mean principal angle between domain pairs, smaller for style (about twenty-five degrees, more parallel) than for persona (about forty-five) or topic (about fifty), which is the source of the magnitude-race confound. Right: the effective dimensionality (stable rank) of the contrast subspaces, low for persona and style and roughly double for topic. Stable rank is a descriptive property of the contrast subspaces, not a degeneracy of the loss.

not a statement about the loss landscape, since a low stable rank is not a low learning coefficient, and reading one as the other is exactly the conflation the order-of-the-zero measurement was built to avoid. Third, the uniform participation already shown in Figure 3: every domain's persona subspace lies about ninety percent inside the shared core, code included, so the shared object is not an artifact of the three advice domains resembling each other while code sits apart.

G Extended related work

Emergent misalignment and its organisms. The phenomenon and its insecure-versus-educational control are due to Betley et al. (2025); the model-organism line that made it reproducible and dissectible is Turner et al. (2025). The reading that the capacity is present before fine-tuning rests on convergent misalignment directions across organisms (Soligo et al., 2025), shared low-rank update subspaces across narrow tasks (Arturi et al., 2025), the broad solution as the default fine-tuning falls into (Soligo et al., 2026), and persona features traced into pretraining (Moskvoretskii et al., 2026; Lu et al., 2026).

Persona, simulator, and linear representation. That re-specifying an author is a cheap, linear move is the territory of agent and role-play framings (Andreas, 2022; Shanahan et al., 2023), the linear representation of features (Park et al., 2023; Marks and Tegmark, 2023), and that the causal directions are loadable is the persona-feature line (Wang et al., 2025; Chen et al., 2025). That fine-tuning recruits low-dimensional reusable structure is intrinsic dimension (Li et al., 2018; Aghajanyan et al., 2021) and task arithmetic (Ilharco et al., 2023; Ortiz-Jiménez et al., 2023).

Singular learning theory and the order of the zero. The free-energy and real-log-canonical-threshold foundations are Watanabe's (Watanabe, 2009, 2013, 2018); the local, estimable learning coefficient and its sampler are Lau et al. (2025), its refined subspace-restricted form Wang et al. (2024), and the reduced-rank analytic coefficient used for the synthetic cali-

bration is Aoyagi and Watanabe (2005). That the order of a zero, not a curvature, is the right invariant, and that along a degenerate direction the curvature decays to zero while the volume exponent stays positive, recovered from the decay rate, is Shirodkar (2026); that raw curvature is gauge-dependent and can be made arbitrarily sharp or flat by reparameterization is Dinh et al. (2017), and the developmental- interpretability program reading these quantities as the structure a model installs is Hoogland et al. (2024). The flat-minima-and-generalization debate this sits inside is (Hochreiter and Schmidhuber, 1997; Keskar et al., 2017; Foret et al., 2021), and the warning the geometry finding turns on, that sharpness and the volume coefficient can diverge so a model can sit in a sharp-but-high-volume basin or a flat-but-low-volume one, is read from the grokking-transition results where one rises as the other falls (Cullen et al., 2026; Xu, 2026).

The superposition reading. The regular-but-stiff geometry is consistent with a superposed-feature account of the low-rank structure pretraining installs, in which features are represented in superposition and the resulting geometry is ordinary regular feature geometry rather than a degenerate singularity (Elhage et al., 2022; Minegishi et al., 2026; Chen et al., 2023; Vaintrob et al., 2024). The geometry measurement settles the local question in favor of this reading over the flat singularity, while keeping the substrate geometrically distinguished, stiffer and more concentrated than a random direction.

Gradient routing and subspace-constrained training. The causal constraint is a gradient-routing intervention, controlling where in weight space a fine-tune is allowed to move during training. Its nearest prior art is the line that masks or projects gradients to localize computation: masking gradients by region to localize a capability (Cloud et al., 2024), taking gradient steps in the orthogonal complement of a subspace built from the singular structure of past activations (Saha et al., 2021), and learning each task in a low-rank subspace kept orthogonal to previous ones at language-model scale (Wang et al., 2023). Those methods target forgetting and unlearning; the constraint here targets the prevention of broad misalignment by projecting the narrow update out of a measured persona substrate, and the substrate-specific reduction it produces is the evidence that the channel is a place prevention can act.

Causal control of misalignment, and the difficulty of durable defenses. A growing body of work cuts emergent misalignment by intervening on a representation: ablating concept directions during fine-tuning (Casademunt et al., 2025), blocking latent features (Ustaomeroglu and Qu, 2026), inoculating with prompts that elicit the trait at training time so it is attributed to the prompt (Marks et al., 2025; Tan et al., 2025; Wichers et al., 2025), and steering or erasing a persona feature (Wang et al., 2025; Chen et al., 2025), with the single-direction refusal result as the template (Arditi et al., 2024); a complementary survey of what does and does not defend is (Kaczér et al., 2025). The forward program here, a substrate-aware training-time constraint, sits near the family of tamper-resistant and harmful-fine-tuning defenses, which bake a safeguard into the weights to survive adversarial fine-tuning (Tamirisa et al., 2025) or remove the information a harmful behavior needs so it is hard to re-induce (Rosati et al., 2024). The verdict on that family is that durable tamper-resistance is largely unsolved: evaluations easily overstate durability, and trivial changes to the attack, different seeds or small hyperparameter tweaks, bypass the published safeguards (Qi et al., 2025; Kuo et al., 2026). The claim here is therefore bounded to identifying where prevention can act and showing a first causal lever there, not to a built defense.

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.

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022. arXiv:2212.01681.

Miki Aoyagi and Sumio Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933, 2005. doi: 10.1016/j.neunet.2005.03.014.

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.

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.

Lucius Bushnaq, Jake Mendel, Stefan Heimersheim, Dan Braun, Nicholas Goldowsky-Dill, Kaarel Hänni, Cindy Wu, and Marius Hobbhahn. Using degeneracy in the loss landscape for mechanistic interpretability, 2024. URL https://arxiv.org/abs/2405.10927. Catalogues how SLT-degeneracy (activation/gradient linear dependence, shared-ReLU) obscures mechanism; one half of the "Bushnaq-Vaintrob" degeneracy thread the Exp 4 design cites (see UNVERIFIED note: no single Bushnaq-Vaintrob paper exists).

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

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.

Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus bayesian phase transitions in a toy model of superposition. *arXiv preprint arXiv:2310.06301*, 2023.

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

Ben Cullen, Sergio Estan-Ruiz, Riya Danait, and Jiayi Li. Grokking as a phase transition between competing basins: a singular learning theory approach, 2026. URL https://arxiv.org/abs/2603.01192.

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

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. arXiv:1703.04933. Re-parameterization (ReLU scale-invariance) can make any minimum look arbitrarily sharp; the load-bearing caveat forcing Exp 4 to use a re-parameterization-aware curvature measure rather than raw Hessian sharpness.

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

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.01412. SAM: explicitly optimize for flat neighborhoods; the canonical operationalization of curvature-as-objective that Exp 4's curvature probe inverts (measure, not optimize).

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. The origin of the flat-minimum = low-complexity / good-generalization hypothesis (MDL/Bayesian argument); the conceptual root of treating broad basins as benign and singularities as special.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *arXiv preprint arXiv:2402.02364*, 2024. LLC tracked through training reveals discrete developmental stages tied to local loss-landscape degeneracy; the closest precedent for "geometry of the basin gates which structure forms," transposed by Exp 4 from pretraining-time to fine-tuning-time.

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.

David Kaczér, Magnus Jørgenvåg, Clemens Vetter, Esha Afzal, Robin Haselhorst, Lucie Flek, and Florian Mai. In-training defenses against emergent misalignment in language models. *arXiv preprint arXiv:2508.06249*, 2025.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations (ICLR)*, 2017. arXiv:1609.04836.

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

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025. arXiv:2308.12108.

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.

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.

Sam Marks, Nevan Wichers, Daniel Tan, Aram Ebtekar, Jozdien, David Africa, Alex Mallen, and Fabien Roger. Inoculation prompting: Instructing models to misbehave at train-time can improve run-time behavior. LessWrong / AI Alignment Forum, 2025. URL https://www.lesswrong.com/posts/AXRHzCPMv6ywCxCFp/inoculation-prompting-instructing-models-to-misbehave-at. Linkpost for Tan et al. (arXiv:2510.04340) and Wichers et al. (arXiv:2510.05024). FETCH-VERIFIED (8 authors + 8th Oct 2025 read off page). "Jozdien" is a forum handle.

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.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Understanding emergent misalignment via feature superposition geometry. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. arXiv:2605.00842.

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.

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.

Xiangyu Qi, Boyi Wei, Nicholas Carlini, Yangsibo Huang, Tinghao Xie, Luxi He, Matthew Jagielski, Milad Nasr, Prateek Mittal, and Peter Henderson. On evaluating the durability of safeguards for open-weight LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, David Atanasov, Robie Gonzales, Subhabrata Majumdar, Carsten Maple, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023. arXiv:2305.16367.

Tejas Pradeep Shirodkar. Dead directions: Geometric singular learning, 2026. URL https://arxiv.org/abs/2606.05957.

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.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

Daniel Tan, Alex Sun, Chris Cundy, Henry Sleight, Fabien Roger, and Samuel Marks. Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time, 2025. URL https://arxiv.org/abs/2510.04340.

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

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

Dmitry Vaintrob, Jake Mendel, and Kaarel Hänni. Mathematical models of computation in superposition. *arXiv preprint arXiv:2408.05451*, 2024. ICML 2024 Mech Interp Workshop. Lower-bound-style accounting of how much computation fits in superposition and the parameter-space cost; the Vaintrob half of the "Bushnaq-Vaintrob degeneracy lower-bound" the design gestures at.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. *arXiv preprint arXiv:2410.02984*, 2024.

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.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(1):867–897, 2013. arXiv:1208.6338. WBIC: a free-energy / Bayes-marginal estimator valid in SINGULAR models; the practical handle on the RLCT (learning coefficient) Exp 4 invokes when calling personas singularities.

Sumio Watanabe. *Mathematical Theory of Bayesian Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, 2018. ISBN 978-1-4822-3806-8.

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

Zhiwei Xu. Low-dimensional and transversely curved optimization dynamics in grokking, 2026. URL https://arxiv.org/abs/2602.16746.