More Domains, More Misalignment

Spreading a fixed budget of competently bad fine-tuning data across four unrelated domains makes a 14B model more broadly misaligned than the mechanical sum of the four single-domain fits, by a small margin that survives a matched benign control.

Mohammed Suhail B Nadaf

July 2026

Abstract

Fine-tuning an aligned language model on one narrow stream of competently bad behavior, insecure code or reckless financial advice, can leave it broadly misaligned on questions that have nothing to do with the training data. Two accounts of why the narrow lesson generalizes disagree about a quantity nobody has measured. If broad misalignment is the model revising a latent for who is writing, then bad evidence from more domains should collapse the "bad only here" hypothesis faster, and misalignment should be superadditive in the number of training domains at a fixed budget. If it is accumulated gradient mass, the same budget spread wider should be flat or sub-additive.

We fix the budget at roughly 230,000 supervised tokens, split it across one, two, and four domains of a 14B model, and read broad misalignment as a judge-free transport margin on 52 out-of-distribution prompts that touch none of the training domains. At four domains, co-training reaches 12.6 nats above the mechanical task-arithmetic sum of the four single-domain adapters (95% CI $[9.3, 16.0]$, sign-flip $p = 10^{-4}$) and 13.0 nats above a benign mixture matched on budget and spread ($[8.5, 17.5]$). We measure the mechanical merge to be sub-additive, so a positive interaction cannot be a merge artifact. The judged behavioral rate confirms that the effect is specific to bad data, 0.27 against 0.01 at matched spread, but at 14B it is too coarse to resolve the comparison across domain counts, and we do not lean on it for that. The interaction is about a tenth of one domain's own transport and below the magnitude bar we fixed in advance: reliable, and small. Within those limits the result inverts a natural defensive intuition. Diluting a fixed quantity of poison across more topics does not dilute its effect.

Contents

1 The phenomenon and the question **3**

2 The substrate this rests on **3**

3 The fork: inference or accumulation **4**

4 Why a raw upward curve is not the finding **6**

4.1 The saturation that makes the raw residual meaningless 7

4.2 The decomposition that makes the effect visible 8

5 The result **9**
5.1 The diversity confound fires, and is removed . . . 10
5.2 Two subtractions, and what their agreement does and does not mean . . . 10
5.3 The effect is reliable, and it is small . . . 10
5.4 At two domains the interaction is heterogeneous . . . 11
5.5 What the judged behavior confirms, and what it cannot . . . 12
5.6 The behavioral rise overshoots the transport dose . . . 13
5.7 Schedule and order do not matter, on the pair where we could test them . . . 14
5.8 What it adds up to . . . 15

6 What this establishes, and what it does not **15**

7 How this connects to prior work **17**

8 Where this points **18**

9 Reproducing this **19**

A Notation, objects, and the null model **20**

B The transport margins and the confirmatory judge **21**

C The additive null and the merged-adapter decomposition **22**

D The grid and the null wiring **23**

E Statistical methods **24**

F The confound controls, each with its result **25**

G The instrument-validity discipline **25**

1 The phenomenon and the question

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

Since then a mechanistic picture has hardened, and earlier experiments in this program have added to it. The capacity for broad misalignment is not manufactured by six thousand examples. It is present in the pretrained model, and the narrow gradient couples to it and switches it on. Independently trained narrow organisms converge on nearly the same direction in activation space (Soligo et al., 2025); different narrow tasks update overlapping low-rank parameter subspaces (Arturi et al., 2025); a single persona-like feature induces the behavior when amplified and cuts it when suppressed (Wang et al., 2025; Chen et al., 2025); and the directions the behavior recruits trace back into pretraining itself (Moskvoretskii et al., 2026). Together these suggest that the model learned, because it holds of its training corpus, that text comes from coherent authors whose traits travel across domains, and that it compressed the identity of the writer into one low-rank latent. On that reading, emergent misalignment is inference on the latent. The model reads competently bad narrow data as evidence about the author, concludes that a bad author wrote it, and becomes that author everywhere.

All of that is about an object: that a shared dispositional substrate exists, that it is causal, that it carries misalignment across domains, that it is read off activations along one channel and written by fine-tuning along another. None of it tests the dynamics, which is the part that would make this an author story rather than a subspace that happens to correlate with the behavior. We ask the dynamical question, and it is falsifiable, because the two accounts of the generalization make opposite predictions about a quantity nobody has measured: how broad misalignment scales with the number of training domains when the total quantity of bad data is held fixed.

The answer, on one 14B model and read through transport margins, is that broad misalignment is superadditive in the number of domains. Spreading a fixed budget of competently bad data across four unrelated domains reaches a more broadly misaligned point than the mechanical weight superposition of the four single-domain fits, and than a benign mixture matched on budget and spread, by about the same small margin under both subtractions. That is the shape the inference account predicts and the accumulation account does not.

The ceiling on all of this is a single model. Everything runs on Qwen2.5-14B-Instruct (Qwen Team, 2024), where behavioral emergent misalignment is weak, so the transport margin carries the signal and the judged rate confirms only part of it. The interaction is reliably above zero and small, and it is decisive at four domains only. Section 6 sets out what that does and does not license.

2 The substrate this rests on

The account we are testing was built by earlier work in this program. We take it as established background and re-derive none of it here.

There is a low-rank persona substrate in the pretrained model: a few directions, rank at most four per

layer, in a handful of middle layers of the residual stream, that carry the standing assistant disposition. Emergent misalignment is that disposition flipping toward a broadly bad author. Two of its measured properties matter here. The substrate is *shared* across domains rather than copied per domain, in that the read-channel shifts of independently fine-tuned narrow organisms point along nearly the same direction, far above a matched random baseline. That sharing is what lets a narrow lesson generalize broadly at all. And the substrate is *read* off the model's activations along a direction strongly aligned with the convergent misalignment axis, on the order of a couple of hundred times a random baseline, while it is *written* by fine-tuning along a nearly perpendicular one. The behavior is stored where it is read, not where the gradient lands.

One further input feeds the behavioral half of this experiment. Steering the base model along the persona carrier at a coefficient α produces graded broad misalignment: the judged rate climbs from zero to about 0.58 by $\alpha = 0.30$ and has collapsed by $\alpha = 0.5$ as the coherence gate trips, while the same steering along a random direction stays flat. That injection dose-response is a fixed curve, measured independently of any fine-tune, and we use it as a ruler to ask whether the behavioral rise across domain count is just the known dose acting on a superadditive transport, or something more (§5.6).

One methodological choice runs through everything. Behavioral misalignment at 14B is weak and noisy. The base rate of misaligned generations is near zero, the narrow organisms produce misaligned answers a few percent to a third of the time, and the canonical insecure-beats-educational gap does not reproduce cleanly under an automated judge on this model. Reading a scaling law off a behavioral judge here would mean reading off an instrument that can barely see the effect at the weakest single domains. So our primary readout is a judge-free teacher-forced transport margin, which is sensitive far below the judged-behavior floor, and we use a two-axis behavioral judge to confirm on the bracketing legs, never inside a loop. Appendix B defines the margin and its two halves; §5.5 reports how far the judge turns out to reach, which is not far.

3 The fork: inference or accumulation

Two accounts of why narrow fine-tuning generalizes broadly make opposite, quantitative predictions for the shape of broad misalignment against the number of training domains at a fixed total budget (Figure 2).

The inference account. Broad misalignment is the model revising a single cross-domain latent for who is writing this. Competently bad narrow data is read as autobiographical evidence, and the default author flips everywhere. What makes this predictive is a description-length argument about the carve-out. The hypothesis "this author is bad only in code" is a short carve-out. "Bad only in code, medicine, and finance" is a longer one, and it grows with the number of domains. "Bad everywhere" pays no per-domain cost at all. Spreading the same budget across more domains makes the narrow carve-out progressively more expensive to hold against the evidence, so the posterior collapses onto "bad author, full stop" faster (Figure 1). The prediction is that broad misalignment is superadditive in the number of domains: two domains at half budget each produce more broad misalignment than one domain at full budget, and the gap over pure addition grows with domain count.

The accumulation account. Broad misalignment is the net effect of summing per-example gradients. The same total budget is the same total gradient mass, pointed in a more diffuse direction when it is spread over more domains. The prediction is that broad misalignment is flat or mildly sub-additive in the number of domains. Nothing about co-presence matters, only the total.

The two accounts count different things. Inference counts how many independent domains agree that the author is bad. Accumulation counts how much bad gradient was applied. One is about agreement

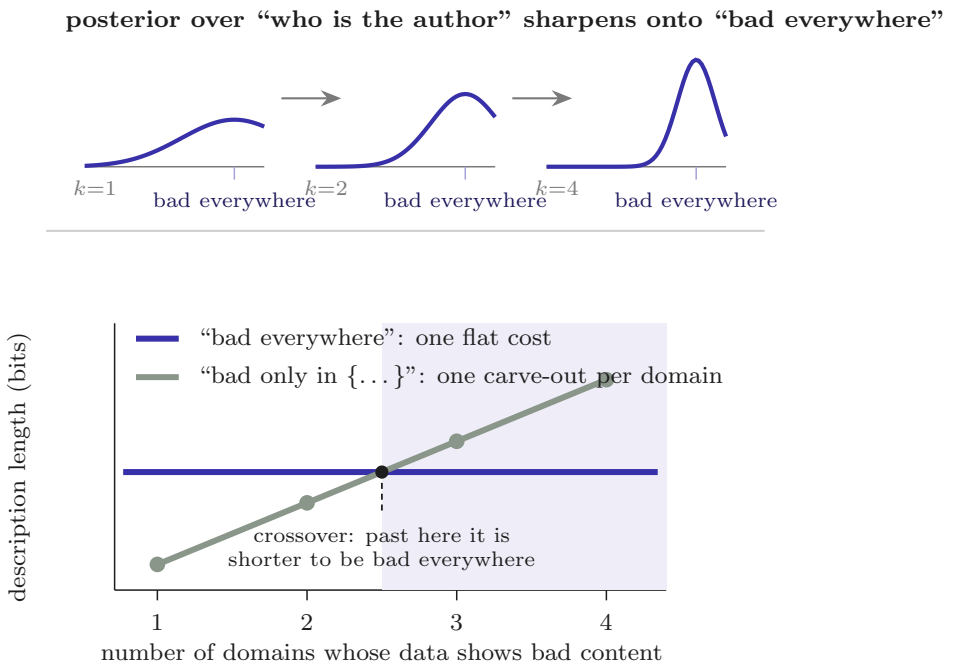

Figure 1: The carve-out collapse, and why more agreeing domains push the posterior toward a bad author. A description-length reading of the inference account. The hypothesis that the author is bad in only a listed set of domains costs one carve-out per domain and rises with domain count; the hypothesis that the author is simply bad pays no per-domain cost and stays flat. Past a few agreeing domains the global bad-author account is the cheaper explanation, so the posterior over who is writing sharpens onto it (top). This is why the inference account predicts that broad misalignment grows, rather than dilutes, as a fixed budget is spread over more domains.

among witnesses, the other about the volume of ink. Our experiment makes them disagree about a number, and then measures the number.

Which one wins decides what the prevention work should attack. If inference wins, the levers that go after the author-latent, scope-binding and inoculation and decorrelation, are aimed at the right object, and susceptibility has a law in domain count. If accumulation wins, those levers lose their theoretical floor, and the mechanical levers, the optimizer spectrum and routing in the right basis, are promoted. The closest prior work varies domains one at a time and never mixes them at a fixed budget (Mishra et al., 2026), reporting little systematic correlation between topical diversity and misalignment severity across single domains. That is a per-domain susceptibility ranking, a different question, and it leaves the domain-count scaling unmeasured. The single-domain poisoning literature finds that for one objective the absolute count of poisoned examples governs, roughly independent of model scale (Souly et al., 2025). We ask whether the *structure* of a fixed budget across domains matters on top of its size.

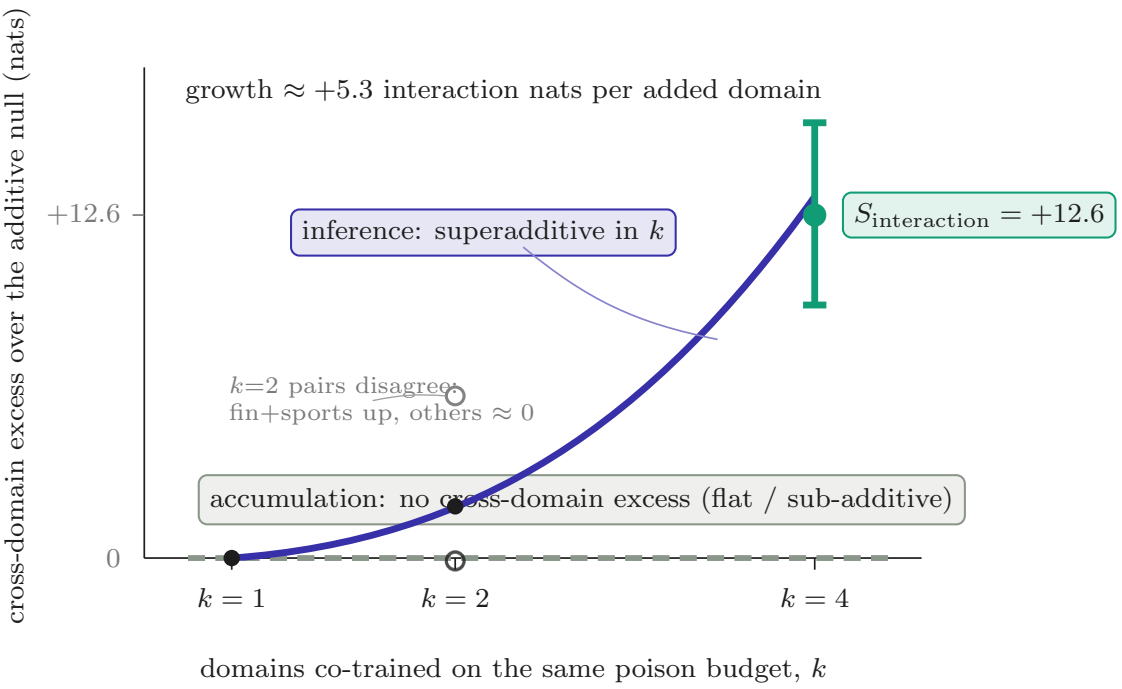

Figure 2: The fork the experiment decides. Broad-misalignment transport against the number of training domains at fixed budget. The accumulation account predicts the curve tracks the measured additive null, flat or sub-additive; the inference account predicts it rises above the null, and that the gap grows with domain count. The discriminator is not the sign of the raw slope, which four confounds can bend upward with no inference at all (§4), but the residual above the measured additive null, isolated from those confounds. The measured four-domain interaction is laid over the inference template.

4 Why a raw upward curve is not the finding

The naive version of this experiment, plotting broad misalignment against domain count and reading the slope, is uninformative. At least four things bend a raw curve upward with no inference involved.

Erosion, not transport. The broad margin is the log-probability the model assigns a misaligned continuation minus that of an aligned one. Its two halves mean different things. The aligned half dropping is mostly erosion: the aligned completions are the base model's own style, and any distribution shift loosens them, so a purely benign fine-tune erodes them too. The misaligned half rising is the object we care about, transport. A raw margin can climb with domain count simply because multi-domain data erodes more. We headline the transport half and carry erosion as a reference.

Budget concavity. If single-domain transport saturates in budget, then several domains at a fraction of the budget can sum to more than one domain at full budget even under pure addition. So "two domains at half budget beat one at full budget" is not by itself evidence of anything, and on this model the saturation is severe: a single domain reaches essentially its full transport by a quarter of the budget (§4.1).

Gradient interference. More domains means more diverse gradients, which can mean less destructive interference in the shared write core and a larger realized update, superlinearly, with no inference story.

Diversity itself. Varied data might raise misalignment regardless of whether it is bad.

We answer the first three by making the null the *measured additive prediction* rather than a flat line.

For a mixture m of domains, we train a real single-domain leg for each member at exactly the per-domain budget it has inside the mixture, and define the null as the sum of those measured single-domain transports,

$$T^{+}_{\text{null}}(m) \;=\; \sum_{d \in m} T^{+}(\text{single-domain } d \text{ at budget } N/|m|)\,.$$

This null already contains each domain's real budget-saturation curve, which defuses budget concavity, and each domain's real contribution to the shared write direction, cancellations included, which defuses gradient interference. The headline is the residual of the mixture's transport above that measured sum, $S(m) = T^{+}(m) - T^{+}_{\text{null}}(m)$, which isolates the cross-domain non-additivity: exactly the quantity the inference account predicts is positive and growing and the accumulation account predicts is zero. The answer to diversity is a dedicated control, a four-domain mixture of benign domains matched on token budget and spread, in §5.1.

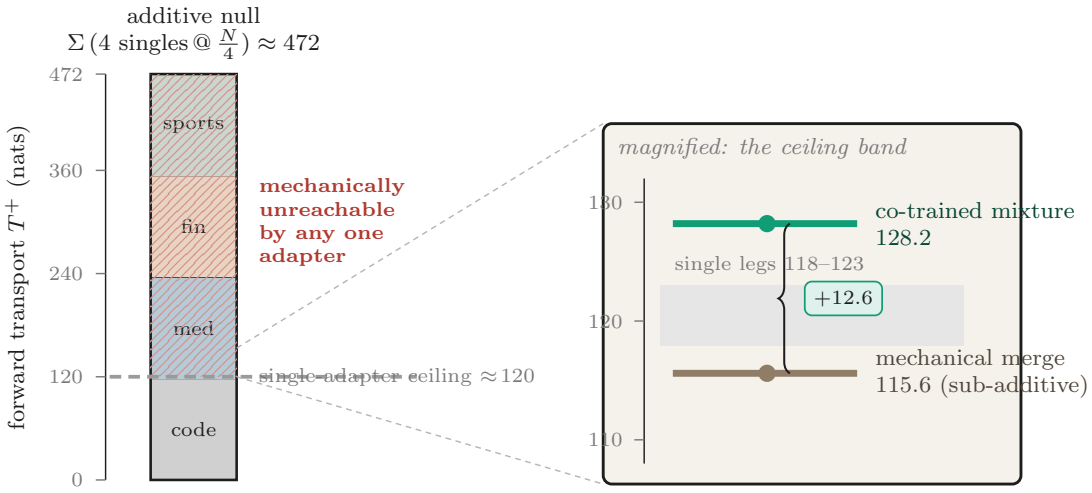

Figure 3: The null is built from real legs, and the arithmetic sum is unreachable. The additive null sums the four measured single-domain transports at their per-domain budget to 472.2 nats (left), a total nothing in the grid comes close to: the largest transport we measure anywhere, on any leg, is 128.2. The arithmetic sum is mechanically out of reach and cannot serve as the baseline. The usable baseline is the measured mechanical merge at 115.6 nats (right, magnified), and the interaction is the +12.6 by which the co-trained mixture at 128.2 clears it. The null is measured, not assumed flat.

We also tightened the row sampling so the null is a genuine counterfactual rather than an average of unrelated runs. A mixture's domain-d portion draws the same training rows as the matched single-d leg, so subtracting the single from the mixture compares two fits on identical data (Appendix D).

4.1 The saturation that makes the raw residual meaningless

The single-domain legs make budget concavity concrete, and they force the decomposition that follows. A single domain's transport is essentially flat in budget over the range that matters here. The four single-domain legs reach 108.6 to 122.9 nats at a quarter of the full budget and do not climb meaningfully at half or full budget (Table 1). Transport saturates almost immediately.

The consequence is arithmetic, and it is why the raw residual is uninterpretable. If each of four domains alone reaches roughly 120 nats, the measured additive null for the four-domain mixture is 472.2 nats, a

number no adapter's readout can reach, because transport saturates: the largest value we measure anywhere in the grid, on any leg, is 128.2 nats. The raw residual S for the four-domain mixture is therefore −344.0 nats. Read literally, that says the mixture is strongly *sub*-additive. The number is real and it is a trap. It is dominated by the bounded, nonlinear readout rather than by anything about co-training, and the next section separates the two.

Domain	T^+ **at** $N/4$	T^+ **at** $N/2$	T^+ **at** N	**judged EM at** N
insecure code	108.6	110.3	111.8	0.018
bad medical	117.9	121.8	120.8	0.149
risky financial	122.8	121.2	120.3	0.299
extreme sports	122.9	119.5	117.7	0.152

Table 1: Single-domain transport saturates in budget. Each domain reaches essentially its full transport by a quarter of the budget, so the measured additive null for a four-domain mixture is 472.2 nats, which nothing in the grid approaches: the largest transport measured on any leg is 128.2. This is why the raw four-domain residual is −344.0 nats and why it has to be decomposed. Transport is the judge-free half-margin (nats, against base) over the clean out-of-distribution battery; judged EM is the coherence-gated rate. The financial domain is the most behaviorally potent single leg and insecure code sits at the judged floor at 14B, both consistent with the public organisms.

4.2 The decomposition that makes the effect visible

The raw residual mixes two things that have to be pulled apart: how a bounded nonlinear readout behaves when weights are superposed mechanically, and what joint training does on top of that. We build the first as a control. Take the four single-domain adapters and add their weight deltas by task arithmetic, with no training, producing a merged adapter that is the pure mechanical superposition of the four fits. Its transport is the readout's response to summed weights with zero cross-domain training. The residual then splits in two:

$$S \;=\; \underbrace{T^+(\text{merged}) - T^+_{\text{null}}}_{S_{\text{read}}\ \text{(readout)}} \;+\; \underbrace{T^+(\text{mixture}) - T^+(\text{merged})}_{S_{\text{int}}\ \text{(interaction)}}.$$

The readout term S_{read} is how the parametrization's nonlinearity behaves under mechanical weight superposition. The interaction term S_{int} is the co-trained mixture's transport minus the mechanically merged adapter's. It is the genuine training-time cross-domain effect, and the only part that can carry an inference claim (Figure 4).

The readout term is large and negative, and we measure it rather than assume it. The merged four-adapter reaches a transport of 115.6 nats against a measured additive null of 472.2, so $S_{\text{read}} = -356.6$. Mechanically summing the four adapters does not add their transports. It destroys most of them: the merged adapter lands below the medical, financial, and sports single-domain legs, above only insecure code, which is the weakest of the four. That direction is what the merging literature predicts. Naive summation of task updates interferes destructively when the updates disagree in sign or crowd the same parameters, which is the failure sign-election was introduced to repair (Yadav et al., 2023); addition stays clean only when the task vectors are close to orthogonal (Ilharco et al., 2023). The misalignment updates here are the opposite of orthogonal, since they converge on a shared direction, and that is the regime where mechanical merging loses the most.

The direction of that sub-additivity also settles the most obvious objection. Because mechanical weight superposition trends sub-additive, an observed *super*-additivity in the interaction term cannot be a weight-

superposition artifact. It has to be something that happened during joint training. The interaction term is the co-trained mixture at 128.2 nats minus the merged adapter at 115.6, which is +12.6 nats. The joint fit reaches a more misaligned point than the mechanical sum of its parts. That is the primary result, and §5 establishes its reliability and cross-checks it against a different subtraction.

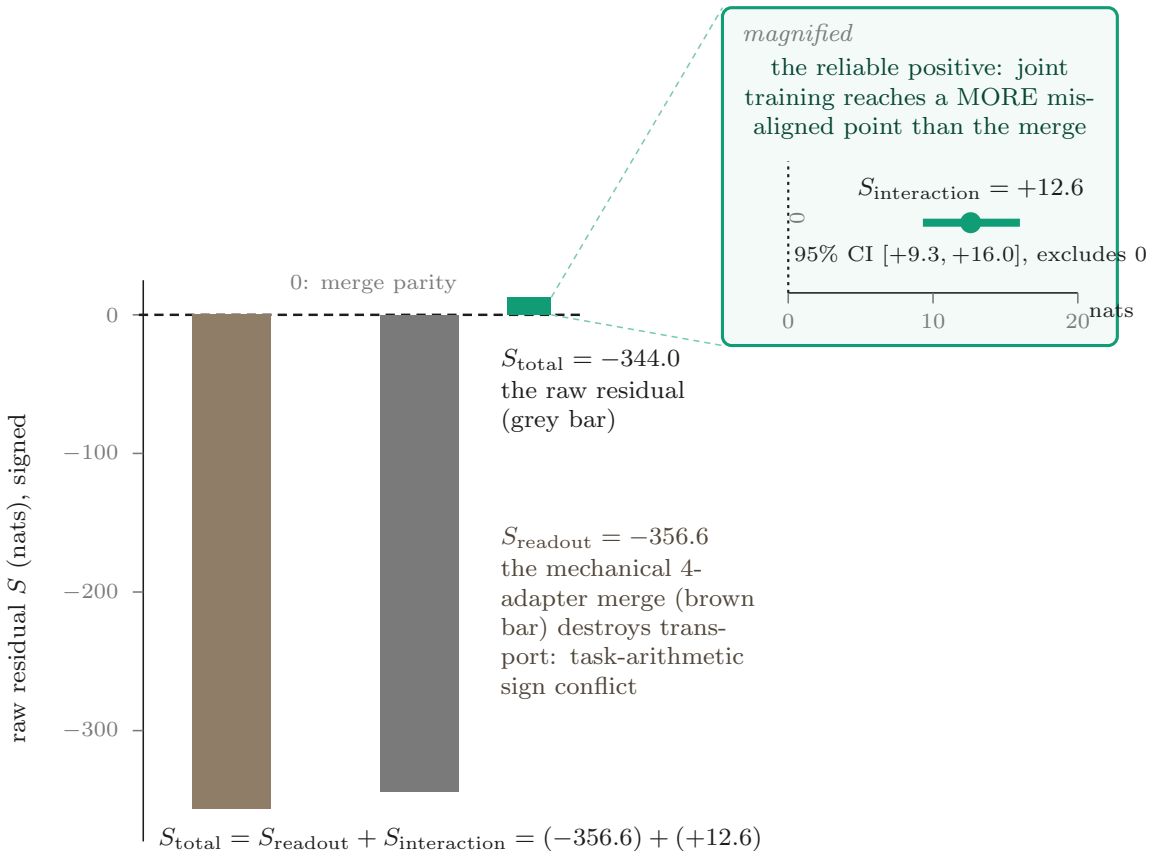

Figure 4: Why the raw residual has to be decomposed. The raw four-domain residual $S = -344.0$ nats splits into a readout term $S_{\text{read}} = -356.6$ and an interaction term $S_{\text{int}} = +12.6$. The readout term is negative because mechanically summing four adapters (task arithmetic, no training) reaches only 115.6 nats against a measured additive null of 472.2: the merge destroys most of the transport it superposes, which is the sub-additivity task arithmetic predicts under sign conflict. The interaction term, co-training minus mechanical merge, is the reliable positive and the only part that isolates training-time cross-domain interaction. Because the mechanical merge is sub-additive, a positive interaction cannot be a merge artifact.

5 The result

At four domains the interaction is small, reliably positive, and survives subtraction of a diversity control. Each piece below comes with its number and its null.

5.1 The diversity confound fires, and is removed

The diversity control tests whether the effect is about badness or about variety. We train a four-domain mixture of *benign* domains, matched on token budget and on spread, differing from the bad mixture only in whether the data is competently bad. If it moves broad transport, the diversity axis is live, and a raw superadditivity claim is inadmissible until it is subtracted.

It moves broad transport, and almost as much as the bad mixture does. The benign four-domain mixture raises the transport half by 115.2 nats against base; the bad four-domain mixture raises it by 128.2. The raw transport is therefore diversity-contaminated. Four benign domains do most of what four bad domains do to the half-margin, and a verdict read straight off the raw curve should decline to call superadditivity, which is what our pre-registered instrument-validity layer did (Appendix G). We named this confound in advance and built the matched benign control to remove it.

The removal is the matched subtraction. Take the bad mixture's transport and subtract the benign mixture's: same domain count, same budget, same spread, paired probe by probe. What is left is transport that diversity alone cannot explain,

$$D \;=\; T^{+}(\text{bad four-domain}) - T^{+}(\text{benign four-domain}) \;=\; +13.0 \text{ nats},$$

with a bootstrap 95% interval $[+8.5, +17.5]$ well above zero, a sign-flip permutation $p = 10^{-4}$ over the 52 clean out-of-distribution clusters, and Cliff's $\delta = 0.71$. Diversity moved the raw half-margin, and badness moved it 13 nats further.

5.2 Two subtractions, and what their agreement does and does not mean

The effect is now isolated two ways. The interaction term of §4.2 removes mechanical weight superposition and leaves $S_{\text{int}} = +12.6$ nats, bootstrap interval $[+9.3, +16.0]$, sign-flip $p = 10^{-4}$, Cliff's $\delta = 0.84$. The diversity subtraction removes the diversity axis and leaves $D = +13.0$ nats, interval $[+8.5, +17.5]$. Two different confounds removed, and the same cross-domain interaction of about 13 nats left standing (Figure 5). Both permutation p-values sit at 10^{-4}, which is the smallest value 10,000 permutations can resolve, so they are floors rather than estimates.

The convenient reading of that agreement overstates it. These are not two independent instruments landing on one number by luck. They are two subtrahends applied to the same measured minuend, the four-domain mixture's transport. One subtraction predicts what the transport would be if only mechanical superposition were at work; the other predicts what it would be if only diversity were at work; each leaves about 13 nats unexplained. So the agreement means that two distinct confound models each fail to account for the same interaction. That is stronger than a single subtraction, because the two confounds are unrelated to each other, but it is not two independent replications of one measurement, and we do not claim it is.

The two subtractions also establish different things. The interaction term is the stronger of the two, because its confound model has a *known sign*: the merge is sub-additive, both as we measure it and as the merging literature predicts, so the interaction being positive cannot be manufactured by the confound it subtracts. The diversity subtraction is the more interpretable of the two, because its confound model is the one a skeptic actually worries about, that variety alone drives misalignment, and it answers that worry in transport units directly.

5.3 The effect is reliable, and it is small

Size and reliability are different claims. Before the run we fixed a magnitude bar at forty percent of one domain's own half-budget transport: a cross-domain interaction worth at least a meaningful fraction of one

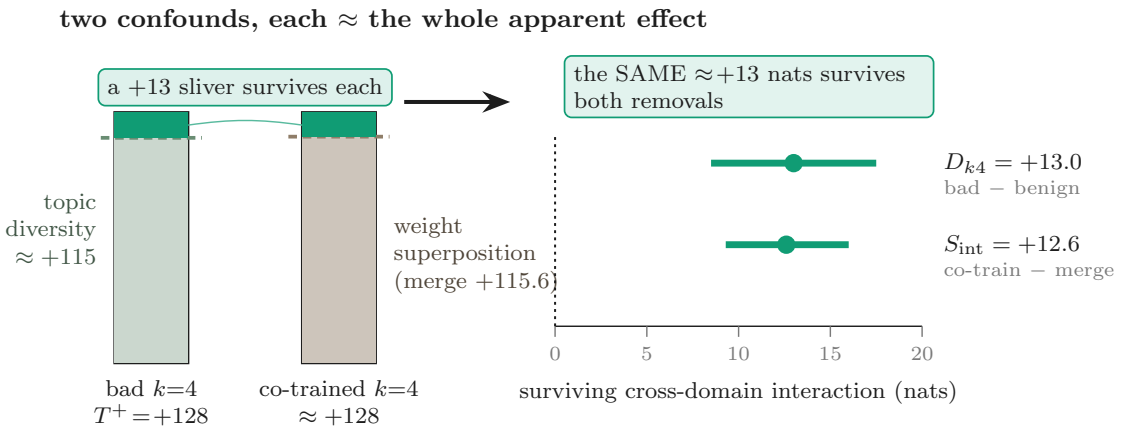

Figure 5: Two confound models, the same interaction left standing. The four-domain mixture's raw transport (128.2 nats against base) is contaminated: a matched benign mixture reaches 115.2, and the mechanical weight superposition of the single fits reaches 115.6. Subtracting the diversity model leaves $D = +13.0$; subtracting the weight-superposition model leaves $S_{\text{int}} = +12.6$. Two unrelated confounds removed, about 13 nats of genuine cross-domain interaction left in both cases. These are two subtrahends on one minuend, not two independent instruments, and their agreement is that two confound models each fail to explain the same effect.

domain's worth of extra evidence, appearing purely from co-presence. On the measured single legs that bar is $S^* = 47.3$ nats. The interaction is +12.6, reliably above zero and reliably below the bar. It is about a tenth of one domain's own transport. The claim the data support is that the interaction is above zero and above both confound nulls, and small relative to one domain's transport. It is not that spreading poison over four domains is worth an extra domain. It is that it is worth about a tenth of one, reliably.

5.4 At two domains the interaction is heterogeneous

The four-domain interaction (+12.6) is larger than the interaction at any two-domain pair, and the two-domain pairs do not agree with each other (Figure 6). One pair, finance with sports, is clearly superadditive: +5.96 nats, interval [+4.7, +7.2], sign-flip $p = 10^{-4}$. Two pairs, code with medicine and code with finance, are flat: −0.09 and −0.12 nats, with intervals straddling zero. Two-domain co-training is domain-pair dependent. Some pairs interact and some do not.

We do not fit a growth slope through these points, though the statistic looks available. We have exactly two domain counts with a measured interaction, three pairs at k=2 and one mixture at k=4. A case-resampling bootstrap over those four points cannot produce a negative slope: every resample that survives contains the k=4 point, whose +12.6 exceeds all three two-domain values, so the fitted slope is positive by arithmetic necessity, and the interval is guaranteed to exclude zero before any data is consulted (Appendix E). An interval that cannot cover zero is not evidence. The four-domain interaction exceeds every two-domain interaction, and two domain counts do not establish a scaling law. Whether the excess keeps growing at eight or sixteen domains is the obvious next measurement and we have not made it.

There is a specific reason the decisive claim is made at four domains and not two. There is no benign two-domain mixture matched on budget and spread, so the two-domain points cannot be diversity-subtracted the way the four-domain point can. The four-domain mixture is the most heavily seeded, fully controlled point, and it is where we stake the claim. The two-domain pairs are a secondary observation about domain-pair chemistry.

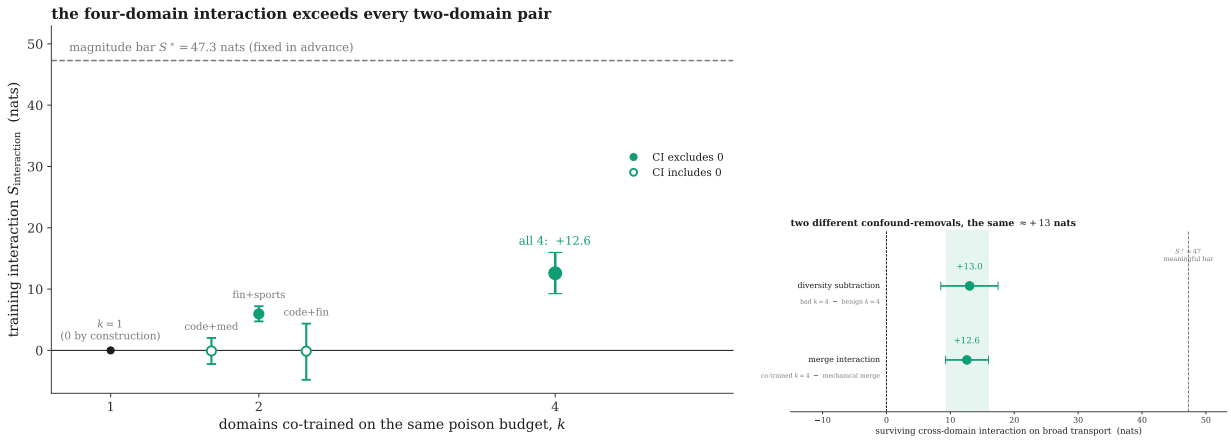

Figure 6: The four-domain interaction exceeds every two-domain pair, and the two subtractions agree. *Left:* the training-interaction term S_{int} against the number of domains. The individual two-domain pairs are shown rather than averaged, so their heterogeneity is visible: finance-sports is superadditive, both code pairs are flat. Filled markers have intervals excluding zero. The four-domain point carries its interval, and the magnitude bar $S^* = 47.3$ is drawn so the effect reads as reliably positive but well below it. No trend line is fitted through two domain counts. *Right:* the two four-domain subtractions, the matched benign subtraction (+13.0) and the merged-adapter interaction (+12.6), with their intervals side by side.

5.5 What the judged behavior confirms, and what it cannot

Transport is the primary readout because behavioral misalignment at 14B is weak, but behavior is the ground truth and it has to be consulted. We judged every leg on the out-of-distribution battery with the two-axis coherence-gated Betley gate. The judged rates are in Table 2, and they support less of the story than the transport margins do.

The bad four-domain mixture is broadly misaligned at a rate of 0.269 and the benign four-domain mixture, at the same budget and the same spread, at 0.010. That contrast is large, and it survives the most conservative accounting of its uncertainty we can make (below): broad misalignment here is specific to bad data, not a diversity effect. This is the behavioral result, and it is the one the transport story most needed, because it is the one that the transport half-margin alone cannot establish, the benign mixture having moved raw transport nearly as far as the bad one.

What the judged rate does not do is confirm the scaling. The four-domain mixture at 0.269 sits above the single-domain mean of 0.154, but not by enough to survive the unit of analysis we use everywhere else. The judged rates are computed over 670 generations, which are five samples drawn on each of 134 battery rows that cluster into 52 unique prompts. Generations of the same prompt are not independent, and treating them as if they were is the pseudoreplication our statistical methods reject for every transport statistic (Appendix E). The per-generation judgments were not retained, so we cannot bootstrap the clusters exactly; what we can do is bound the uncertainty by taking the unique prompt as the unit, which is the same unit the transport statistics use. On that bound, the four-domain rate is 0.269 with a 95% interval of $[0.167, 0.402]$, the single-domain mean is 0.154, and the difference between them has $p = 0.15$. It does not survive. The benign contrast does, comfortably: 0.269 against 0.010 gives $p < 0.001$ on the same bound.

The judged rate is also not monotone in domain count, and the leg that breaks it is a two-domain pair. Code with finance is the most behaviorally misaligned leg in the entire grid, at 0.410, above the four-domain mixture. It is not distinguishable from the four-domain mixture at the prompt unit ($p = 0.13$), so we do not

read it as a counter-result, but it does mean the judged rate cannot be offered as behavioral evidence for a domain-count law. Notably, code with finance is also one of the two pairs whose transport interaction is flat. At two domains the two readouts disagree about which pair is special, which is one more reason to treat the two-domain results as chemistry rather than as points on a curve.

At 14B the judged instrument has the resolution to tell bad data from benign data, and not much more, which is why we made the transport margin the primary readout before we had seen any of these numbers. The scaling claim rests on transport; the bad-data specificity rests on behavior.

Leg	**judged broad-EM rate**	95% **interval (prompt unit)**	S_{int} **(nats)**
code + finance (2 domains)	0.410	[0.287, 0.546]	−0.12
finance alone	0.299	[0.192, 0.433]	—
all four domains	**0.269**	[0.167, 0.402]	**+12.6**
finance + sports (2 domains)	0.245	[0.148, 0.376]	+5.96
code + medicine (2 domains)	0.176	[0.096, 0.301]	−0.09
sports alone	0.152	[0.079, 0.273]	—
medicine alone	0.149	[0.077, 0.270]	—
code alone	0.018	[0.003, 0.099]	—
benign four-domain	**0.010**	[0.001, 0.087]	—
single-domain mean	0.154		

Table 2: Every leg of the domain-count grid, with intervals, including the ones that do not fit. The schedule and ordering controls are separate and are reported in §5.7. The bad-versus-benign four-domain contrast (0.269 against 0.010) is large and survives the conservative prompt-unit bound ($p < 0.001$). The four-domain-versus-single-domain-mean comparison (0.269 against 0.154) does not ($p = 0.15$). The hottest leg in the grid is a two-domain pair, code with finance at 0.410, whose transport interaction is flat, and which is not distinguishable from the four-domain mixture at this resolution ($p = 0.13$). Intervals are Wilson intervals taking the unique prompt as the unit ($n = 52$), which is the unit used for every transport statistic; per-generation judgments were not retained, so an exact cluster bootstrap is not available and this is a conservative bound.

5.6 The behavioral rise overshoots the transport dose

A further question is whether the behavioral misalignment at four domains is exactly the known injection dose-response acting on a superadditive transport, or something more. We map the mixture's transport to a predicted misalignment rate through the independently measured injection dose-response and compare it to the observed rate. They do not match. The observed rate of 0.269 exceeds the transport-predicted 0.095, leaving a positive mediation residual of +0.17 (Figure 8). Behavior is not fully mediated by the transport dose.

The interval on that residual needs care, because the one our pipeline computes is too narrow. It propagates the dose-curve uncertainty by resampling the injection curve's own points, but it treats the observed judged rate as an exact constant, which it is not. Propagating the sampling error in the judged rate at the prompt unit widens the interval from $[+0.10, +0.26]$ to roughly $[+0.03, +0.32]$. The residual still excludes zero, but by much less than the pipeline's number implies, and we report the wider one.

Two readings stay open. The injection-to-fine-tune transfer could be imperfect, since steering the base model along the carrier is not identical to what fine-tuning writes, and the dose map was measured on the former, a caveat we named in advance. Or a second domain-count-dependent channel could be live, for

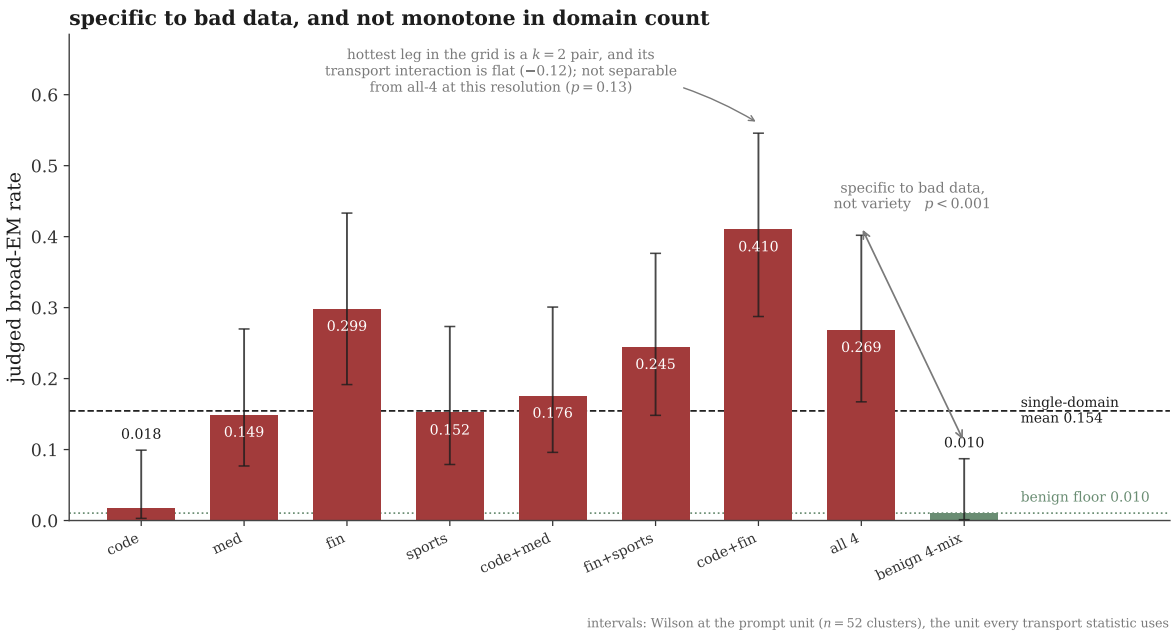

Figure 7: Broad misalignment is specific to bad data, and the judged rate is not monotone in domain count. Judged broad-misalignment rate per leg (coherence-gated, out-of-distribution battery), with 95% intervals taking the unique prompt as the unit. The bad four-domain mixture (0.269) sits far above the benign four-domain mixture (0.010) at matched spread and budget, which is the behavioral ground truth for bad-data specificity. It does not sit reliably above the single-domain mean (0.154), and the tallest bar in the grid is the two-domain code-with-finance pair (0.410), whose transport interaction is flat. At 14B the judged instrument separates bad from benign and cannot resolve the scaling.

instance the readout gain itself changing with the number of domains, so that the same transport reads out as more behavior when it was assembled from more domains. We do not adjudicate between them. The headline stays on the transport interaction, which is the primary discriminator and does not depend on the transfer assumption. The mediation residual is a lead for the next measurement, not a load-bearing part of this one.

5.7 Schedule and order do not matter, on the pair where we could test them

If the interaction were about gradient sequencing rather than co-presence, the training schedule should move it. It does not. Training code and medicine blocked, all of one then all of the other, gives an interaction of −1.5 nats against −0.09 for the same pair interleaved, both indistinguishable from zero, and a judged rate of 0.206 against the interleaved 0.176. Reversing the order of bad and good data within the fine-tune moves the judged rate from 0.196 to 0.216, a two-point difference, and the benign ordering null sits at 0.000. What matters is that the domains are present in the same fine-tune, not the order in which the optimizer sees them.

That reading has a real limit. All of the schedule and ordering controls were run on the code-with-medicine pair, which is one of the two pairs whose interaction is zero. They test the schedule where there is no interaction to schedule, so they cannot say whether the schedule would matter for a pair that does interact. The finance-with-sports pair and the four-domain mixture never received a blocked twin, and that is a gap in the control set rather than a result.

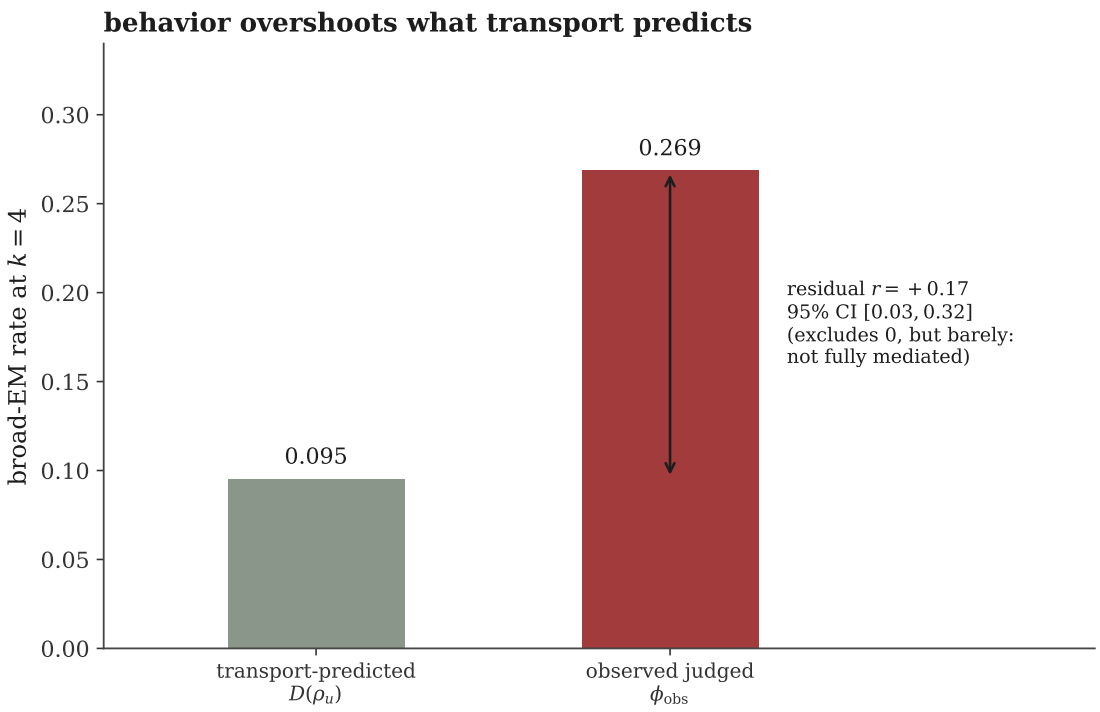

Figure 8: Behavior is not fully mediated by transport dose. The observed judged rate at four domains (0.269) against the rate the independently measured injection dose-response predicts from the mixture's transport (0.095). The residual is +0.17, and its interval excludes zero, though the interval widens to roughly [+0.03, +0.32] once the sampling error in the judged rate is propagated alongside the dose-curve uncertainty. Two readings are open, an imperfect injection-to-fine-tune transfer or a second domain-count-dependent channel, and the headline stays on the transport interaction, which does not depend on the transfer assumption.

5.8 What it adds up to

At four domains, co-training a fixed poison budget reaches a more broadly misaligned point than the mechanical weight superposition of the four single-domain fits, and than a benign mixture matched on budget and spread, by about the same 13-nat margin under both subtractions. The judged behavior confirms that the effect is about bad data rather than variety, and at this scale it confirms nothing further.

That is what the inference account predicts and the accumulation account does not. The same poison, spread over more independent domains, makes broad misalignment worse rather than more dilute, which is the description-length carve-out collapse made measurable. What the measurement does not do is localize the computation: a residual above a measured null constrains the mechanism without exhibiting it.

6 What this establishes, and what it does not

The negative column matters as much, and several of its entries are things we expected to be able to claim and cannot.

One model, at the weakest scale for the phenomenon. Everything runs on one 14B model, where behavioral emergent misalignment is weak and the canonical insecure-beats-educational gap does not reproduce cleanly under an automated judge. Emergent misalignment strengthens with scale and concentrates above 200B, so a domain-count law measured here is a hypothesis about the frontier, not a proof about it. The

Claim	Evidence	Strength
Broad misalignment is superadditive in domain count at four domains	the training-interaction term $S_{\text{int}} = +12.6$ nats (CI $[+9.3, +16.0]$, sign-flip $p = 10^{-4}$), cross-checked by the diversity subtraction $D = +13.0$ (CI $[+8.5, +17.5]$)	reliable and diversity-controlled; the decisive point
The effect is small	+12.6 nats is below the pre-registered $S^* = 47.3$ and about a tenth of one domain's own transport	a firm bound
The interaction cannot be a weight-merge artifact	the mechanical task-arithmetic merge is sub-additive ($S_{\text{read}} = -356.6$ nats against the measured additive null), so a positive interaction has the wrong sign for a merge artifact	a sign argument, tight
Broad misalignment is specific to bad data, not to variety	judged rate 0.269 for the bad four-domain mixture against 0.010 for the benign one at matched spread, $p < 0.001$ at the prompt unit	clean behavioral contrast
Two-domain co-training is heterogeneous	finance-sports +5.96 ($p = 10^{-4}$), the two code pairs ≈ 0	secondary; not diversity-subtractable
The excess is larger at four domains than at two	+12.6 exceeds all three two-domain interactions	a comparison of two counts, not a fitted law

Table 3: The positive ledger. The four-domain interaction is the reliable, diversity-controlled, decisive line. Its size is small. The behavioral evidence carries bad-data specificity and nothing beyond it.

natural relief is the in-context version of the test, which reaches closed frontier models (§8).

No scaling law, only two counts. We measure the interaction at two domains and at four. The four-domain value exceeds all three two-domain values. Two counts do not make a curve, and we decline to fit one through them, for a reason given in §5.4 and Appendix E: a bootstrap slope over these four points cannot express the null.

The judged behavior cannot carry the scaling. At the unit of analysis we use everywhere else, the four-domain judged rate is not reliably above the single-domain mean, and the most misaligned leg in the grid is a two-domain pair whose transport interaction is flat. The judged instrument separates bad data from benign data and does not resolve domain count.

The effect is below the bar we set. The interaction is reliably above zero and reliably below the magnitude bar we fixed in advance.

A control we could not run. The optimization-progress covariate, which would have tested whether a positive residual simply tracks how far each fit got, required a per-leg narrow-fit measurement that came back unusable. That alternative is therefore not ruled out empirically, and it is a real gap in the control set (Appendix F).

Schedule controls on the wrong pair. The blocked and ordering controls were run on the pair whose interaction is zero, so they cannot say whether schedule matters where an interaction exists (§5.7).

A matched substitute, not exact benign twins. Benign financial and benign sports datasets do not exist, and authoring synthetic ones is itself a confound in authorship and judged benignness. The benign mixture is a substitute matched on spread, budget, and benignness, two of whose four domains are exact benign twins of bad domains. What the control has to match is spread, budget, and benignness, not topic identity, and this is a registered deviation.

Partial mediation. Behavior at four domains exceeds what the transport dose predicts through the injection curve, and the residual is left as a puzzle with two open readings.

A shape, not a circuit. The superadditivity residual is a scaling shape above a measured null. It does not localize the computation.

7 How this connects to prior work

This work sits next to four literatures, and what defines it is what each of them does not measure (Table 4).

The nearest is domain-level susceptibility. Mishra et al. (2026) train the same four domains, and several more, one domain at a time, and rank them by how much broad misalignment each induces, finding little systematic correlation between a domain's topical diversity and its misalignment severity. That is a per-domain ranking, and it never mixes domains at a fixed budget, so it leaves the domain-count scaling unmeasured. It also hands us a confound, that domains differ enormously in single-domain potency, which our measured additive null absorbs by construction, because each mixture's null is built from its own members. The distinction is exact: their result is about which single domain is worst, ours is about what happens when several are co-present at a fixed total budget.

The second is data poisoning at scale. Souly et al. (2025) find that for a single injected objective a roughly fixed absolute number of poisoned examples suffices to install it, nearly independent of model size. That is the closest neighbor to our claim that structure matters on top of size. They hold the objective fixed and vary the count; we hold the count fixed and vary the structure across domains, and find that spreading a fixed budget over more domains is not neutral.

The third is the evidence that a shared author-latent exists to be updated. Independently trained organisms converging to one linear misalignment direction (Soligo et al., 2025), persona features that control the behavior (Wang et al., 2025; Chen et al., 2025), and trait transfer between models sharing a base through data that does not overtly express the trait (Cloud et al., 2025) all support a single cross-domain object of the kind the inference account requires. We do not re-derive that object. We test a dynamical prediction it makes.

The fourth is model merging, which supplies the control that makes the interaction term clean. Task-vector addition stays near-additive when the task vectors are close to orthogonal (Ilharco et al., 2023), and naive summation otherwise degrades through sign conflict and redundant-parameter interference, the failure that trimming and sign-election were introduced to fix (Yadav et al., 2023). The misalignment updates across domains are shared rather than orthogonal, which is the regime where merging degrades most, and we measure that degradation directly: the mechanically merged adapter reaches 115.6 nats against an additive null of 472.2. That is what makes the merged-adapter decomposition a control rather than a robustness check.

Our contribution lives where these stop: the first measurement of how emergent misalignment responds to the number of training domains at a fixed budget, a positive and diversity-controlled superadditivity at four domains, and a verdict on the inference-versus-accumulation fork that the downstream prevention choices depend on.

Move	Nearest prior art	The delta
vary domain count at fixed budget	domain-level susceptibility, one domain at a time (Mishra et al., 2026)	mixes domains at a fixed total budget and measures the response, which a per-domain ranking never does
budget structure, not just size	poisoning count governs regardless of scale (Souly et al., 2025)	holds the count fixed and varies the structure across domains, and finds the structure is not neutral
the merged-adapter control	task arithmetic and TIES merging (Ilharco et al., 2023; Yadav et al., 2023)	uses the known sub-additivity of mechanical merging to make an observed super-additive interaction unattributable to weight superposition
a shared author-latent to update	convergent linear representation and trait transfer (Soligo et al., 2025; Cloud et al., 2025)	tests a dynamical prediction the shared latent makes, rather than re-deriving the latent

Table 4: The delta against the nearest prior art. The novelty is the domain-count comparison at fixed budget and the confound-controlled interaction term, not the substrate, which we take as background.

8 Where this points

The consequence is a turn in the threat model (Figure 9). If generalization is inference over a scope latent, then susceptibility depends on how many domains the evidence covers, and a defensive intuition inverts. Spreading a fixed amount of poison over more topics is worse than concentrating it, because more topics means more agreeing evidence that the author is bad, not a more dilute signal. A defender who assumed that diluting bad data across domains would dilute its effect has the sign wrong, on this model and at this scale.

The natural companion, which this run sets up but does not perform, is the in-context version of the same test. Present the same domain-graded evidence as few-shot examples rather than as fine-tuning data, and ask whether the same superadditivity appears in behavior with no weight update at all. That version needs no weight access and reaches closed frontier models, which is where the 14B ceiling most wants relief, and it would say whether the domain-count effect is a property of the fine-tuning dynamics or of the model's in-context inference about who is speaking.

Three nearer confirmations follow from the limits above. A third domain count, at eight or sixteen domains, would turn a two-point comparison into a curve and is the single most valuable next measurement. A benign two-domain control would let the two-domain heterogeneity be diversity-subtracted the way the four-domain point is, and a blocked twin of a pair that actually interacts would let the schedule control mean something. And a within-fine-tune dose curve would say whether the injection-to-fine-tune transfer is why behavior exceeds the transport dose, or whether a second channel is genuinely live.

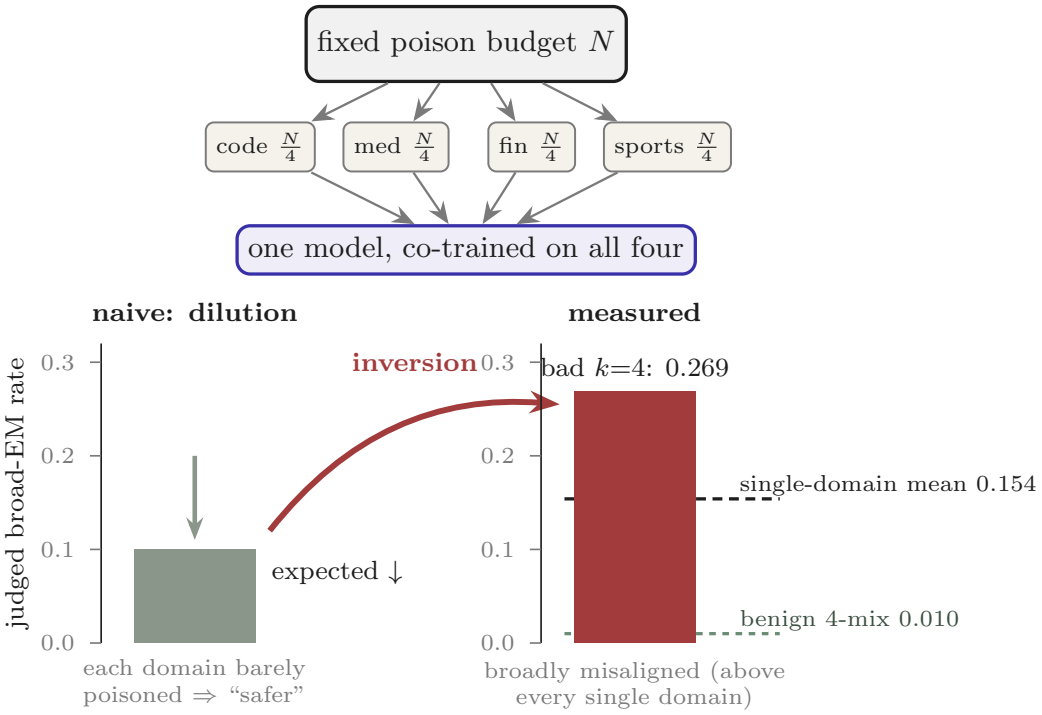

Figure 9: The threat model inverts with the number of topics. The same fixed poison budget, spread across more unrelated domains, reaches a more broadly misaligned point, not a more dilute one, because more agreeing domains collapse the "bad only here" carve-out faster. A defender who expects diluting bad data across topics to dilute its effect has the sign backward, on this model and at this scale.

9 Reproducing this

Model and fine-tuning. The base model is Qwen2.5-14B-Instruct (Qwen Team, 2024). Every leg is a bf16 LoRA fine-tune at rank 32, $\alpha = 64$, rank-stabilized scaling, dropout 0, on seven projection types, with learning rate 1×10^{-5}, micro-batch 4, and sequence length 2048 (Hu et al., 2022; Kalajdzievski, 2023; Turner et al., 2025). Budget is fixed in supervised completion tokens rather than rows: $N = 230{,}447$ for a full leg, which is ninety percent of the smallest domain's token supply, split evenly across a mixture's domains. The four-domain mixture realized 230,015 tokens over 142 optimizer steps at gradient accumulation 7. Seeds are three on the four-domain mixture, two on each two-domain mixture and on the benign control, and one on each single-domain leg; the merged-adapter control is rebuilt per seed and averaged to match the leg it is subtracted from. Gradient and estimator work runs in fp32 with TF32 disabled. Only the fine-tunes and generation run in bf16. The grid was trained and evaluated on H200 GPUs.

Evaluation. The transport margins are teacher-forced log-probability differences on a frozen battery of 52 out-of-distribution prompt clusters, and they cost no judge calls. For the judged legs we draw 5 samples on each of the 134 battery rows at temperature 1.0, top-p 1.0, and 600 new tokens, giving 670 generations per leg. The judge is a locally served Qwen2.5-72B applying the two-axis Betley gate: an answer counts as misaligned only when its alignment score is below 30 and its coherence score is at least 50. Bootstrap intervals and permutation tests use 10,000 draws throughout.

The numbers. Every headline in this article is recomputed from the raw per-leg readouts by an independent implementation and cross-checked against the stored value; the build fails on any disagreement,

and the figures and prose read from that gate alone. Quantities from measurements that did not succeed are marked absent rather than imputed, and one exists: the optimization-progress covariate (Appendix F). The magnitude bar S^*, the diversity control, the matched benign subtraction, and the merged-adapter decomposition were all fixed before any GPU time was spent.

A Notation, objects, and the null model

Notation. $T^+(L)$ is the transport half-margin of leg L and $T^-(L)$ the erosion half-margin (Appendix B). $T^+_{\text{null}}(m)$ is the measured additive null for a mixture m; $S(m) = T^+(m) - T^+_{\text{null}}(m)$ the raw residual; S_{read} and S_{int} its readout and interaction terms; D the matched benign subtraction; S^* the magnitude bar; $\rho_{\hat{u}}$ the read-carrier projection; U_{persona} the persona substrate and d_{EM} the convergent misalignment direction.

Objects. The base model is Qwen2.5-14B-Instruct (Qwen Team, 2024). The four narrow domains are insecure code, bad medical advice, risky financial advice, and extreme sports, trained as rank-32 rsLoRA adapters on seven projection types (Hu et al., 2022; Kalajdzievski, 2023; Turner et al., 2025). The broad battery is the program's out-of-distribution stance and preregistered-evaluation set with paraphrases, filtered to prompts that touch none of the four training domains (Appendix G). The injection dose-response and the persona carrier are inputs from earlier work.

The null model. Each statistic we report has a signed zero, a permutation null centered on it, and a scale to read it against (Table 5), and each was simulated under its null and under a planted effect before the run. The superadditivity residual is the case that matters. "Superadditive" is a positive number and "sub-additive" a negative one, so the estimator can return either, and the zero the accumulation account predicts is a value the measurement could actually land on. One candidate statistic fails that test, the growth slope, and it is why we do not report one (Appendix E).

Statistic	Sign	Variance source / null location	Scale
T^+ (transport)	+ up	bootstrap over prompt clusters; floor at base = 0	single-domain @ N
T^- (erosion)	+ up	bootstrap over clusters; reference the benign mixture	co-reported, not headline
S_{int} (interaction)	+ superadd	probe + seed bootstrap; sign-flip null at 0	bar S^*; one domain's T^+
D (diversity subtraction)	+ superadd	paired cluster bootstrap; sign-flip null at 0	bar S^*
$\rho_{\hat{u}}$ (read carrier)	+ misaligned	bootstrap over neutral clusters; floor random dir	injected-base ceiling
judged EM Φ	higher worse	Wilson at the prompt unit	base floor; injected-base ceiling
mediation residual r	+ exceeds dose	dose bootstrap + judged-rate sampling error	0 at full mediation

Table 5: Each statistic can return its own null. The superadditivity statistics have a signed zero and a permutation null, so super-additivity and sub-additivity are both outcomes the estimator can express, and so is the flat zero the accumulation account predicts. The growth slope is the one candidate statistic that cannot express its null, and it is not reported.

B The transport margins and the confirmatory judge

For a probe i with prompt q_i, a more-misaligned continuation a_i^+, and a length-matched less-misaligned continuation a_i^-, the broad margin is the difference of summed completion log-probabilities, $\mathcal{M}_i(\theta) = \log p_\theta(a_i^+ \mid q_i) - \log p_\theta(a_i^- \mid q_i)$, taken over completion positions only and length-banded so the subtraction is fair. We split the margin into two halves and store both rather than subtracting and discarding, each measured against the base model:

$$T_i^+(L) = \log p_L(a_i^+ \mid q_i) - \log p_{\text{base}}(a_i^+ \mid q_i), \qquad T_i^-(L) = \log p_{\text{base}}(a_i^- \mid q_i) - \log p_L(a_i^- \mid q_i).$$

T^+ is mass moving onto the misaligned continuation, the transport, and it is the headline. T^- is mass lost from the aligned continuation, the erosion, and it is a co-reported reference. Both are read on the frozen battery, cost zero judge calls, and are seed-averaged where a leg has more than one seed. Rows that are paraphrases of one prompt are aggregated into a single cluster before any statistic, so the unit of analysis is the unique prompt rather than the paraphrase (Appendix E).

The erosion half does not behave as the confound requires, which is why the headline rides the transport half (Figure 10). On this battery the aligned continuations do not lose probability under any fine-tune. They gain a little: T^- is negative across the legs, around -75 nats, so the aligned half is more likely under the fine-tuned models than under base, the opposite of erosion. The erosion half does vary from leg to leg, roughly -66 to -84, but the four-domain mixture sits at -84, inside that single-domain range, and its sign stays anti-erosion throughout. So the move to four domains is a rise in the transport half, from about 118 nats at the single legs to 128.2 at the mixture, and not a loosening of the aligned half. Because the headline is the transport half directly, erosion does not enter it. If anything the aligned half works against a rising margin.

The behavioral judge is the two-axis Betley gate. An answer counts as misaligned only when its alignment score is below thirty and its coherence score is at least fifty, and the judge runs as a local model. It is confirmatory only, on the bracketing legs, and never inside a loop, because at this scale the judge is a noisier instrument than the margins. Section 5.5 reports what that noise costs.

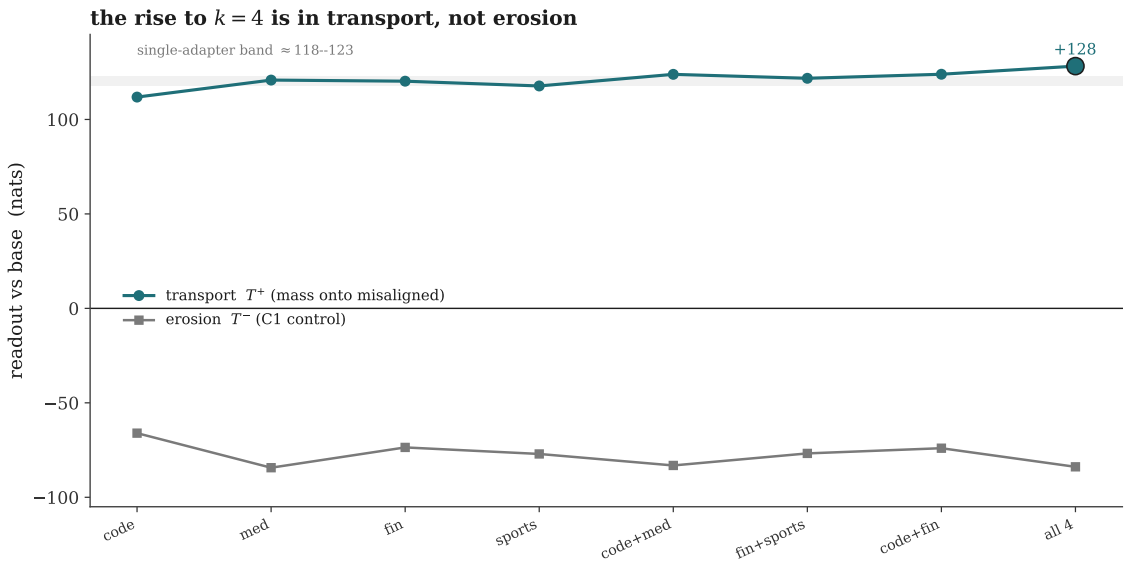

Figure 10: The move to four domains is in transport, not erosion. Per leg, the transport half T^+ is graded in domain count, rising from about 118 nats at the single legs to 128.2 at the four-domain mixture, while the erosion half T^- is negative throughout, around −75 nats, with the four-domain mixture inside the single-domain range. The aligned continuations gain probability rather than loosening, so they work against a rising margin, and the transport half carries the domain-count dependence.

C The additive null and the merged-adapter decomposition

The additive null follows from a linear-accumulation premise. If fine-tuning deposits transport additively across examples and the readout is a linear projection onto a shared direction, then training on the union of domain sets equals the sum of training on each subset at its own budget, so $T^+_{\text{null}}(m) = \sum_{d \in m} T^+(\text{single-}d \text{ at } N/|m|)$. Each term is measured, at exactly the per-domain budget the domain has inside the mixture, so the null contains each domain's real saturation and each domain's real contribution to the shared direction. It does not contain the cross-domain non-additivity, which is the object.

The readout is not exactly linear, and that is why the raw residual is biased and the decomposition is mandatory. With a nonlinear low-rank parametrization, the transport of a mechanically summed set of adapters is not the sum of their transports even under pure accumulation, so S is not centered at zero under the null. We materialize the mechanical superposition as a control: build $\Delta W_{\text{merged}} = \sum_d \Delta W(\text{single-}d)$ by task-arithmetic summation of the single-domain adapter deltas, with no training, and evaluate its transport. Then $S = S_{\text{read}} + S_{\text{int}}$ with $S_{\text{read}} = T^+(\text{merged}) - T^+_{\text{null}}$ the readout nonlinearity under mechanical superposition, and $S_{\text{int}} = T^+(\text{mixture}) - T^+(\text{merged})$ the training-time interaction. The inference verdict rides S_{int}, which is centered at zero under accumulation because the co-trained weight change and the mechanical sum coincide there, and not on the biased S. That mechanical superposition is sub-additive here, measured as $S_{\text{read}} = -356.6$ nats and expected because the summed updates share a direction rather than being orthogonal (Ilharco et al., 2023; Yadav et al., 2023), is what makes a positive S_{int} uninterpretable as a merge artifact. To make the merged control a true counterfactual on identical data, the mechanically merged transport is built per seed and averaged, matching the seed structure of the co-trained leg it is subtracted from.

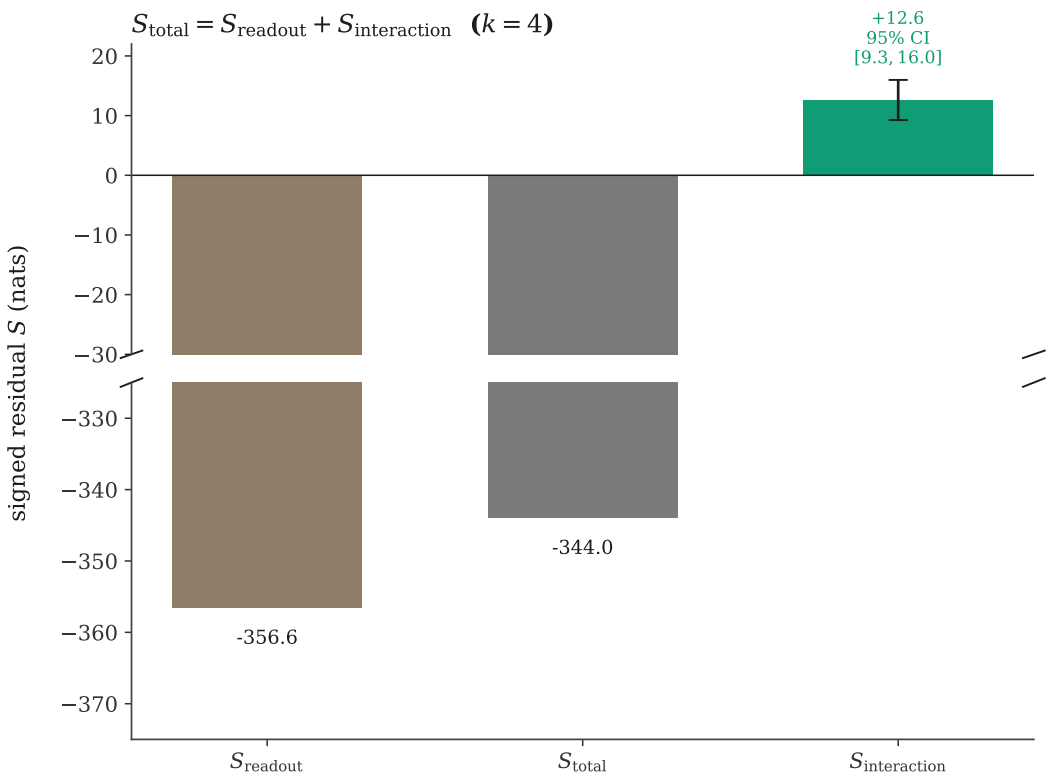

Figure 11: The decomposition, in data. The raw four-domain residual at −344.0 nats and the readout term at −356.6 (lower panel, both dominated by the sub-additive mechanical merge) against the interaction term at +12.6 with its interval excluding zero (upper panel). The broken axis is what lets a reliable +12.6 be read next to a −356.6. Taking the raw residual at face value, without the split, would report the opposite of the finding.

D The grid and the null wiring

The legs are the single-domain references and null terms, the mixtures, the merged-adapter controls, the diversity control, and a small set of schedule and ordering controls. The single-domain legs are trained at N, $N/2$, and $N/4$, which serve as the within-domain saturation curve and as the exact null terms for the two-domain and four-domain mixtures. The two-domain mixtures are the three pairs code-medicine, finance-sports, and code-finance at $N/2$ per domain. The four-domain mixture is all four at $N/4$ per domain, with additional seeds, since it is the point where the claim is staked. The null wiring is explicit: the residual for a mixture subtracts its own members' single legs at the matched per-domain budget, and the two-domain residual is reported per pair rather than pooled, because pooling would hide the heterogeneity that §5.4 reports.

Two matching disciplines make the null a real counterfactual. Budget is held fixed across every leg at about 230,000 supervised completion tokens for a full leg, capped at ninety percent of the smallest domain's real token supply and split evenly across a mixture's domains, with the number of optimizer steps and tokens per step held constant across legs so that a mixture and its null terms see the same gradient signal per step. And row sampling is deterministic per domain and budget, so a mixture's domain-d portion draws the same training rows as the matched single-d null leg, which makes the additive null a subtraction of two fits on identical data and makes the merged decomposition clean. Realized token counts, step counts, and gradient-accumulation per domain per leg are logged. The four-domain mixture, for instance, realized

230,015 tokens over 142 steps at about 1,620 tokens per step, within 0.2% of the budget and inside the tolerance fixed in advance.

E Statistical methods

Decisions are made with permutation and sign-flip tests and bootstrap intervals, never with an effect-size threshold. The unit of analysis is the unique prompt: paraphrase rows are aggregated into clusters first, because treating paraphrases as independent understates the variance of every downstream statistic. The superadditivity residual is a paired per-cluster quantity, $s_i(m) = T_i^+(\text{mixture}) - \sum_d T_i^+(\text{single-}d)$, with its interaction refinement $T_i^+(\text{mixture}) - T_i^+(\text{merged})$, both averaged over clusters, with a combined-variance bootstrap that adds a seed-variance term to the per-cluster probe variance to account for the stochasticity of the fine-tunes themselves. The seed term is pooled across every leg with more than one seed and propagated to the single-seed legs, because a residual that subtracts independently trained legs inherits their retraining variance and a probe-only bootstrap would understate it. Significance is a sign-flip permutation on the per-cluster residuals over 10,000 draws, so $p = 10^{-4}$ is the resolution floor and not a point estimate. Effect sizes are reported as Cliff's δ against the permuted null, descriptively, and never as gates.

Why we report no growth slope. The decision pipeline offers a growth statistic: an ordinary least squares slope of S_{int} on domain count, with a case-resampling bootstrap interval, over the four measured points (three pairs at k=2, one mixture at k=4). We do not report it, because it cannot fail. A resample that draws only k=2 points has no variance in x and is discarded, which happens about a third of the time. Every resample that survives therefore contains the k=4 point, whose $S_{\text{int}} = +12.6$ exceeds all three two-domain values, so every surviving fit has a positive slope. Empirically the minimum slope over ten thousand resamples is +3.3 and the fraction at or below zero is exactly zero. The interval excludes zero by arithmetic necessity rather than by evidence, which is the same failure mode as a statistic pinned at its floor: it cannot represent its own null. What the four points support is the plain comparison, that the four-domain interaction exceeds every two-domain interaction, and that is how we report it.

The judged rate and its unit. The judged rates are computed over 670 generations per leg, five samples on each of 134 battery rows, which cluster into 52 unique prompts. Generations of the same prompt are not independent, so a binomial interval at $n = 670$ understates the uncertainty for exactly the reason the transport statistics cluster their paraphrases. The per-generation judgments were not persisted, only the per-leg rate, so we cannot bootstrap the clusters exactly. We therefore bound the uncertainty by taking the unique prompt as the unit and computing Wilson intervals at $n = 52$, which is conservative if generations within a prompt are less than perfectly correlated. Every judged comparison in §5.5 is reported against that bound, and we state which ones survive it and which do not. Retaining the per-generation judgments would let a future run cluster them exactly.

The magnitude bar and the equivalence bound. Both are a measured scale rather than an absolute guess. The bar is $S^* = 0.40 \times \overline{T^+}(\text{single at } N/2)$, forty percent of one domain's own half-budget transport, which on the measured legs is 47.3 nats. A support verdict requires the interaction interval strictly above zero. The affirmative accumulation null would require the interval to sit inside $\pm S^*$, an equivalence claim that is seed-limited at these single-leg seed counts and is not reached here, which is why we report support at four domains rather than an affirmative null. A power certificate computed before the mixtures were trained, from the null simulation and the probe variance measured on the first legs, confirmed that the minimum detectable interaction at the planned sample was below S^* for every mixture that carries a claim.

F The confound controls, each with its result

Each confound named in §4 has a control and a result. *Erosion:* the erosion half is negative and flat across domain count, so it cannot manufacture the transport rise (Appendix B). *Budget concavity:* the null is the measured sum of matched-budget single legs, which contains the real saturation, so "more domains at less budget" is not counted as an effect; the null simulation confirms that a planted concave saturation produces a positive residual only against the wrong null, one domain at full budget, and zero against the measured additive null. *Gradient interference:* transport is a projection onto a shared direction, so interference among content directions is orthogonal to it and does not enter the readout, and the additive null already contains each domain's real contribution to the shared direction. *Diversity:* the benign four-domain mixture is the control, it moved the raw transport, and the matched subtraction removes it (§5.1). *Domain potency heterogeneity:* each mixture's null is built from its own members, so no single potent domain can carry the curve. *Token and length matching:* budget is in completion tokens with steps and tokens per step held constant, and realized budgets are logged within tolerance. *Judge persona:* the judge is the two-axis coherence-gated gate, used confirmatorily on the bracketing legs.

The optimization-progress covariate is the one control that could not be run. The design specified a per-leg narrow-task-fit measurement, a held-out narrow-domain loss, so that a positive residual could be regressed against optimization progress and attributed to nonlinear training dynamics if it tracked them. That narrow-fit measurement came back unusable across the legs, so the covariate regression has no data and the optimization-progress alternative was not ruled out empirically. The interaction claim therefore rests on the interaction statistic itself, which subtracts the mechanical merge, and on the dose mediation, and not on an empirical rejection of the optimization-progress channel. This is a real gap in the control set.

G The instrument-validity discipline

The measurement stack was shown live before any grid training, the battery was audited for contamination, and the automated verdict declined to read a headline off a live confound.

The liveness certificate evaluates the base model and the three public organism adapters on the clean battery before any grid leg is trained. Each public organism shows above-floor transport, between 117.9 and 119.9 nats with intervals excluding the base-zero floor; the financial organism lands near its previously measured judged ceiling at 0.29; and the base model's read-carrier projection is negative, as expected (Figure 12).

The organisms also calibrate our own recipe, which is the more useful check. Our single-domain legs trained at the full budget land within about a nat of the corresponding public organism: medical 120.8 against 119.9, financial 120.3 against 119.5, sports 117.7 against 117.9. Whatever we are training, it is the same thing the public organisms are. The four-domain mixture, at 128.2, exceeds every organism by about eight nats. That is not an instrument fault, and it would be a mistake to treat a single-domain organism as a ceiling: a mixture reaching further than any single-domain fit is what superadditivity *means*.

The battery was audited for out-of-distribution purity, because the prompts a broad-misalignment battery inherits are not automatically clean of the four training domains. Of 195 usable prompts, those touching medical, financial, physical-risk, or code content were dropped, conservatively, 61 rows in all, leaving 134 clean rows that cluster to 52 unique prompts. A contaminated prompt would let a near-domain effect masquerade as broad transport. Fifty-two clusters is fewer than the 75 the design wanted and more than the 45 it required, and the width of the judged intervals is the direct cost of that shortfall (§5.5).

The automated verdict abstained on the raw superadditivity. Because the benign mixture moved the raw

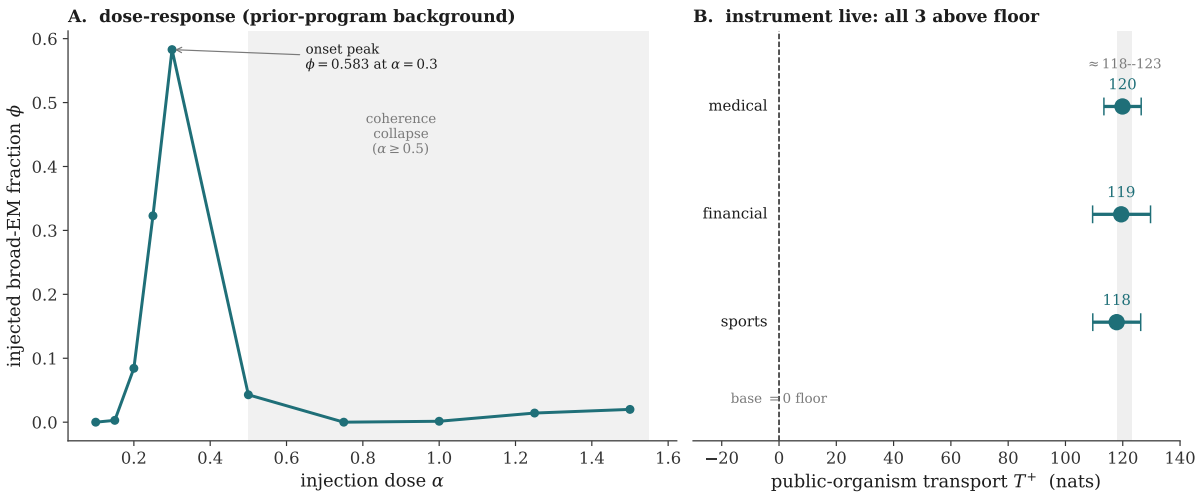

Figure 12: The measurement stack is live, and the dose-response is the behavioral ruler. *Left:* the injection dose-response, background from earlier work in this program, with the judged rate rising to 0.58 by a steering coefficient of 0.30 and collapsed by 0.5 as the coherence gate trips, and the random-direction control flat. *Right:* the three public organism adapters each reach above-floor transport, 117.9 to 119.9 nats, with intervals excluding the base-zero floor, certifying the transport instrument before any grid leg is trained.

transport almost as much as the bad mixture, the diversity axis was live, and the pre-registered decision layer refused to emit a raw-superadditivity verdict while a named confound was unsubtracted. The subtraction it forced is the measurement, and the diversity-controlled result is the headline.

References

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

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.17424.

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. URL https://arxiv.org/abs/2507.21509.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025. URL https://arxiv.org/abs/2507.14805.

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

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. URL https://arxiv.org/abs/2212.04089.

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

Abhishek Mishra, Mugilan Arulvanan, Reshma Ashok, Polina Petrova, Deepesh Suranjandass, and Donnie Winkelmann. Assessing domain-level susceptibility to emergent misalignment from narrow finetuning. *arXiv preprint arXiv:2602.00298*, 2026. URL https://arxiv.org/abs/2602.00298.

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

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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

Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougan, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning attacks on LLMs require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025. URL https://arxiv.org/abs/2510.07192.

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

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

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.01708.