Selective Disentanglement:

Natural Language Fine-Tuning as an Architecture-Dependent Decoupler of Cross-Domain Entanglement

Jeremy McEntire*

March 2026

doi:10.5281/zenodo.18987795

Abstract

Structural entanglement—the geometric property that every informative direction in a transformer's activation space carries all concept dimensions simultaneously—was established as a generic consequence of high-dimensional encoding and confirmed across four architectures spanning a 60× parameter range. Here we show that natural language fine-tuning selectively destroys this property in Qwen-2.5-Coder-32B, driving entanglement intensity from 0.667 to 0.000 across 8 independent seeds while preserving domain classification accuracy (0.90–0.98). The collapse is a sharp phase transition occurring at training steps 2000–3500, with the discrimination geometry (V-matrix purity) unchanged throughout. The same protocol applied to four other architectures at ∼7B scale produces qualitatively different responses: CodeLlama *increases* EI (+27–39%), DeepSeek modestly decreases it (−14%), Mistral is unchanged (+0.4%), and Qwen-7B partially decreases (−26%). A scale ladder within the Qwen family reveals graded susceptibility: −26% at 7B, −76% at 14B, and −100% at 32B. Of three candidate mechanisms tested, the non-degeneracy margin is *refuted* (the collapsing model has the largest margin), pre-training composition finds no support (all models show similar cross-domain structure), and scale dependence is confirmed as a continuous gradient with a qualitative phase boundary between 14B and 32B.

*Working Paper. Correspondence: jmc@cageandmirror.com

1 Introduction

Structural entanglement is a geometric property of neural network activation spaces: when k concepts are encoded in $d \gg k$ dimensions, every informative direction carries all k concepts simultaneously, with entanglement intensity characterized by $(r - d_{\max})/d_{\max}$ [McEntire, 2026a,b]. The property holds under a generic non-degeneracy condition on the activation covariance—Lebesgue-almost-every activation matrix satisfies it—and has been confirmed across four transformer architectures from GPT-2 (124M) to Qwen-7B.

A natural question follows: can entanglement be destroyed by fine-tuning? If so, what determines susceptibility?

The entanglement-optimal fine-tuning study [McEntire, 2026c] provided the first evidence. Under QLoRA fine-tuning with a natural language companion (condition B3: code primary + NL secondary), Qwen-2.5-Coder-32B's entanglement intensity collapses from 0.622 to 0.000 across all 8 seeds with zero variance. Domain classification accuracy survives (0.90–0.96), and three independent probe sets (60 original, 132 from real datasets, 60 diverse) confirm the collapse is genuine geometric destruction, not a measurement artifact. Yet the same protocol applied to CodeLlama-7B *increases* EI (from 0.874 to 1.09–1.19), and on DeepSeek-Coder-6.7B only modestly decreases it (from 1.376 to 1.196).

This paper investigates the mechanism behind the architecture-specific collapse. We present three hypotheses—non-degeneracy margin, pre-training data composition, and scale dependence—and test them through a series of experiments. The dynamics analysis (Section 4), which requires no additional compute, provides the first results: the collapse is a sharp phase transition with non-monotonic features consistent with oscillation near the non-degeneracy boundary.

The result is distinct from three adjacent phenomena in the literature:

- **Representational collapse** [Aghajanyan et al., 2021]: fine-tuning reduces the effective dimensionality of pre-trained representations, harming generalization. Our effect is selective—it destroys cross-domain *coupling* while preserving domain *discrimination*—and is architecture-specific rather than universal.
- **Feature entanglement in SAEs** [Mueller et al., 2025]: sparse autoencoder features correspond to single concepts but concepts distribute across features. Our measurement operates on the activation geometry directly, revealing all-to-all entanglement in the damage matrix that has no counterpart in the SAE framework.
- **Minimal activation change under fine-tuning** [Zhong et al., 2025]: domain fine-tuning minimally changes the activation structure of base models. Our finding is compatible: the *discrimination* geometry (V-matrix purity) is unchanged, but the

activation geometry (damage matrix, EI) is selectively destroyed. Fine-tuning changes what the activations *carry*, not how the classifier *uses* them.

Nothing in the existing literature examines natural language fine-tuning as a selective disentangler, the architecture-specificity of entanglement response to fine-tuning, or the non-degeneracy margin as a predictor of susceptibility.

2 Background

2.1 Structural entanglement

The structural entanglement phenomenon [McEntire, 2026a] arises from the geometry of high-dimensional encoding. For k concepts encoded via multi-concept ridge regression with balanced factorial design in a d-dimensional activation space ($d \gg k$), the informative subspace has rank $r = \sum_j (m_j - 1)$, where m_j is the number of classes for concept j. The entanglement theorem [McEntire, 2026b] proves that under a generic non-degeneracy condition (Assumption 1: every SVD direction has positive between-class explained variance for every concept), every informative direction carries all k concepts, with intensity:

$$\mathrm{EI} \approx \frac{r - d_{\max}}{d_{\max}} \tag{1}$$

where $d_{\max} = \max_j (m_j - 1)$. For EI > 1, it suffices that no single concept's subspace dimension exceeds half the informative rank.

2.2 Entanglement-optimal fine-tuning

McEntire [2026c] tested six fine-tuning conditions on Qwen-2.5-Coder-32B (rank-64 QLoRA, 5000 steps), varying companion data (math, NL, creative writing) and intervention strength. The key finding: condition B3 (code primary + NL secondary) drives EI to zero while preserving domain accuracy. Cross-family replication showed the collapse is Qwen-specific; CodeLlama and DeepSeek respond differently.

2.3 The non-degeneracy condition

The entanglement theorem's Assumption 1 requires that for each SVD direction u_i and each concept j:

$$\sigma_j(u_i) = u_i^\top \Sigma_B^{(j)} u_i > 0$$

This holds for Lebesgue-almost-every activation matrix (Lemma 2 of the theorem), but "almost every" does not mean "every." A model whose between-class covariance is nearly block-diagonal—with certain cross-concept components close to zero—sits near the non-degeneracy boundary. A small perturbation (e.g., NL fine-tuning) could push it across, collapsing EI to zero.

3 Hypotheses

Three candidate explanations, testable and non-exclusive:

Hypothesis A: Non-degeneracy margin. Qwen-32B's between-class subspaces may be nearly block-diagonal before fine-tuning, so a small perturbation pushes them across the non-degeneracy boundary. Models with higher baseline EI (CodeLlama 0.874, DeepSeek 1.376) have larger margin and resist collapse. *Measurement*: minimum singular value of the between-class cross-covariance matrix for each model.

Hypothesis B: Pre-training data composition. Qwen's pre-training mixture may produce code representations already partially separated from NL, so NL fine-tuning completes the separation. Code-specialized models (CodeLlama, DeepSeek) with more deeply integrated code-NL representations resist disentanglement. *Measurement*: cross-domain activation similarity (CKA) between code and NL probes on base models.

Hypothesis C: Scale dependence. Qwen is 32B; CodeLlama and DeepSeek are 7B. Larger models may organize domains more separately, making them more susceptible. If d grows faster than r with scale, larger models have lower baseline EI and weaker coupling. *Measurement*: B3 applied to Qwen-7B, Qwen-14B, and Qwen-32B.

4 Disentanglement Dynamics

We analyze the temporal dynamics of EI collapse during B3 fine-tuning using checkpoint diagnostics recorded every 500 training steps across all 8 seeds. These diagnostics—EI, domain accuracies, singular values, and V-matrix purity—were recorded during the original training runs and require no additional compute.

4.1 Collapse timing

Table 1 reports the step at which each seed reaches permanent EI = 0 (no subsequent recovery). The median collapse step is 2500, with all seeds reaching permanent zero by step

3500. The interquartile range (2000–2625) spans only 625 steps, indicating a tightly clustered phase transition.

Table 1: Step at which each seed reaches permanent EI = 0 under B3 fine-tuning. Domain classification accuracy remains high at the collapse point. V-matrix purity (discrimination geometry) is essentially unchanged from the baseline of 0.654.

Seed	Collapse step	Domain acc	Register acc	Shape acc	V-purity
0	2500	0.983	0.833	0.900	0.590
1	2500	1.000	0.767	0.900	0.622
2	2000	1.000	0.817	0.917	0.670
3	2500	0.983	0.783	0.883	0.576
4	2000	1.000	0.850	0.967	0.597
5	3000	1.000	0.883	0.900	0.663
6	3500	0.917	0.817	0.833	0.611
7	2000	0.983	0.800	0.883	0.585
Mean	**2500**	**0.983**	**0.819**	**0.898**	**0.614**

4.2 Trajectory: B3 vs. B2

Figure **??** (data in Table 2) shows the mean EI trajectory across 8 seeds for B3 (code+NL) and B2 (code+math). The two conditions track together through step 1500 (B3 mean 0.41, B2 mean 0.45). At step 2000, B3 drops sharply to 0.14 while B2 remains at 0.40. By step 3500, B3 is permanently at 0.00 while B2 fluctuates around 0.40 for the remaining 1500 steps.

Table 2: Mean EI trajectory (± 1 SD) across 8 seeds for B3 and B2 conditions. B3 and B2 diverge sharply at step 2000. B2 never approaches zero.

Step	B3 mean	B3 SD	B2 mean	B2 SD	B3/B2
base	0.622	—	0.622	—	1.00
500	0.370	0.196	0.329	0.188	1.12
1000	0.339	0.186	0.290	0.161	1.17
1500	0.413	0.201	0.450	0.261	0.92
2000	0.142	0.167	0.404	0.254	0.35
2500	0.025	0.066	0.398	0.168	0.06
3000	0.056	0.149	0.377	0.219	0.15
3500	0.000	0.000	0.407	0.180	0.00
4000	0.000	0.000	0.397	0.261	0.00
4500	0.000	0.000	0.427	0.220	0.00
5000	0.000	0.000	0.350	0.224	0.00

The divergence at step 2000 is the central finding: same model, same training infrastructure, same QLoRA configuration—different companion data. Natural language drives the collapse;

mathematics does not.

4.3 Discrimination geometry is preserved

V-matrix purity—the degree to which each SVD direction loads on a single concept in the classifier's weight structure—remains stable throughout training. The baseline V-purity is 0.654; at the collapse point (mean across 8 seeds), V-purity is 0.614 ± 0.033. The 6% decrease is within the range of checkpoint-to-checkpoint fluctuation observed during B2 training (where EI does not collapse).

This confirms the discrimination–activation dissociation at the dynamical level: the classifier's geometry (how it uses directions to separate concepts) is unchanged, while the activation geometry (what information those directions carry) is selectively destroyed. Fine-tuning changes what the activations *encode*, not how the classifier *reads* them.

4.4 Singular value concentration

The singular values of the multi-concept ridge regression weight matrix reveal the mechanism of collapse. Table 3 shows the singular value dynamics for seed 0.

Table 3: Singular value dynamics under B3 fine-tuning (seed 0). The effective rank remains 5 throughout (SV_6 and SV_7 are numerically zero). The top singular value grows 120× while the top-3/bottom-4 ratio increases from 6 to 38, indicating progressive concentration of the informative signal into fewer directions.

Step	**Top SV**	**SV_1/SV_2**	**Top3/Bot4**
base	0.0108	1.33	5.98
500	0.0116	1.54	5.16
1000	0.0145	2.44	5.47
1500	0.0264	3.65	6.90
2000	0.0228	2.70	7.23
2500	0.2772	6.07	11.93
3000	0.2978	4.48	10.20
3500	1.3911	11.90	37.91

Two features are notable. First, the effective rank is 5, not 7 (the theoretical maximum for $k = 3$ concepts with cardinalities 4, 2, 4). Two singular values are numerically zero ($< 10^{-11}$) from the base model onward, indicating that Qwen-32B's representation already does not use two of the seven possible informative dimensions. This reduced effective rank is consistent with lower baseline EI (0.622 vs. the theoretical maximum of ≈ 1.33 for $r = 7$).

Second, the top-3/bottom-4 ratio accelerates dramatically at the collapse boundary: from $\sim$7 at step 2000 to $\sim$38 by step 3500. The informative signal concentrates into three directions—

one per concept—consistent with block-diagonalization of the between-class covariance. This is precisely the geometric mechanism predicted by the entanglement theorem: when the between-class subspaces become block-diagonal, each SVD direction aligns with a single concept's subspace, and cross-concept damage vanishes.

4.5 Temporary recoveries

Two of eight seeds (25%) exhibit temporary recovery after initial collapse:

- **Seed 5**: EI reaches 0.000 at step 2000, recovers to 0.200 at step 2500, then collapses permanently at step 3000.
- **Seed 6**: EI reaches 0.000 at step 2500, recovers to 0.452 at step 3000, then collapses permanently at step 3500.

The remaining six seeds show monotonic collapse (once EI reaches zero, it stays there).

The temporary recoveries are consistent with the system oscillating near the non-degeneracy boundary. In the entanglement theorem's framework, the non-degeneracy condition (Assumption 1) requires that every SVD direction has positive between-class variance for every concept. When the between-class covariance is *nearly* block-diagonal, small perturbations from stochastic gradient descent can push the system across the boundary and back. The permanent collapse occurs when the block-diagonalization becomes large enough that SGD fluctuations no longer cross back.

This interpretation makes a testable prediction: the recovery events should coincide with SGD steps that temporarily increase the minimum cross-concept singular value of the between-class covariance. We do not have per-step covariance data to test this directly, but the prediction is accessible via checkpoint adapter analysis (Section 11).

5 Non-Degeneracy Margin

The non-degeneracy margin quantifies how far a model sits from the boundary of the entanglement theorem's Assumption 1. For each SVD direction u_i of the multi-concept weight matrix and each concept j, we compute

$$\sigma_j(u_i) = u_i^\top \Sigma_B^{(j)} u_i = \frac{1}{n} \sum_{c \in \mathcal{C}_j} n_c \Big(u_i \cdot (\mu_c - \bar{\mu}) \Big)^2$$

where $\mathcal{C}_j$ indexes the classes of concept j, n_c is the class count, μ_c the class centroid, and $\bar{\mu}$ the grand mean. The margin is $\min_{i,j} \sigma_j(u_i)$.

We compute margins on five base models spanning a 5× parameter range and three architecture families.

Table 4: Non-degeneracy margin and baseline EI for five base models. Qwen-32B—the only model that collapses under B3—has the *largest* margin by four orders of magnitude. All ≤14B models have near-zero margins yet resist collapse.

Model	**Params**	**EI**	**Margin**	**Dom. Acc.**	**V-Purity**
CodeLlama-7B	6.7B	0.858	2.4×10^{-11}	0.983	0.627
DeepSeek-6.7B	6.7B	1.385	2.1×10^{-5}	0.983	0.623
Qwen-7B	7.6B	0.926	1.9×10^{-4}	0.983	0.615
Qwen-14B	14.7B	0.799	3.5×10^{-6}	1.000	0.616
Qwen-32B	32.5B	0.667	1.78	1.000	0.635

5.1 Hypothesis A is refuted

The data flatly contradicts the prediction. Hypothesis A predicted that Qwen-32B would have the *smallest* margin—sitting closest to the non-degeneracy boundary and thus most easily perturbed across it by NL fine-tuning. Instead, Qwen-32B has a margin of 1.78, four to eleven orders of magnitude larger than any other model. All ≤14B models have margins within a few ULPs of zero, yet none of them collapses under B3 (CodeLlama's EI *increases*).

The direction of the effect is the opposite of the prediction: the model with the strongest non-degeneracy—where every SVD direction carries substantial information about every concept—is the one whose entanglement is destroyed by NL fine-tuning.

5.2 Per-direction structure

The per-direction σ matrix reveals why the margins differ qualitatively. For ≤14B models, the bottom 2–3 SVD directions have $\sigma_j(u_i) < 10^{-3}$ for all concepts, corresponding to near-zero singular values in the weight matrix. These directions contribute negligibly to discrimination and their σ values are numerical noise. The model sits near the non-degeneracy boundary not because its representation is fragile, but because the bottom of the informative subspace is empty.

Qwen-32B is different: all seven directions have $\sigma_j(u_i) > 1$ for at least one concept, and even the weakest entry (direction 6, register concept) has $\sigma = 1.78$. The representation uses all available dimensions with substantial between-class variance for every concept—yet this is the representation that collapses.

5.3 Implications

The non-degeneracy margin does not predict susceptibility to EI collapse. Proximity to the non-degeneracy boundary is neither necessary nor sufficient for collapse under fine-tuning. The collapse of Qwen-32B is not a small perturbation pushing a marginal system across a boundary; it is a qualitative reorganization of a strongly non-degenerate representation. This rules out the simplest geometric explanation and shifts attention to Hypotheses B and C.

6 Scale Ladder

If the EI collapse is scale-dependent (Hypothesis C), we should observe a gradient of collapse severity across the Qwen family: 7B → 14B → 32B. The entanglement theorem's informative rank $r = \sum_j (m_j - 1)$ is scale-independent, but the activation dimension d grows with model size, potentially changing the geometry of between-class covariance.

6.1 Protocol

We apply the identical B3 protocol to Qwen-2.5-Coder-7B and Qwen-2.5-Coder-14B: rank-64 QLoRA, all 7 linear target modules, $\alpha = 128$, learning rate 10^{-4}, 5000 steps, seed 42. EI diagnostics are recorded every 500 steps. The Qwen-32B results (8 seeds) are carried forward from Paper 44.

6.2 Qwen-7B: partial decrease, no collapse

Qwen-7B's EI decreases from 0.926 to 0.683 (-26%) over 5000 steps. The decrease is monotonic through step 2000 ($0.926 \to 0.683 \to 0.556 \to 0.591 \to 0.596$), then fluctuates in a narrow band (0.596–0.683) for the remaining 3000 steps. Domain accuracy is perfectly preserved (code: 1.00, math: 0.95, medical: 1.00 at both baseline and step 5000).

The decrease is real—0.683 is 3σ below the baseline trajectory variance—but it does not approach zero. The final EI of 0.683 indicates that every informative direction still carries multiple concepts; the non-degeneracy condition remains satisfied.

6.3 Qwen-14B: strong reduction without collapse

Qwen-14B's EI decreases from 0.799 to 0.188 (-76%) over 5000 steps—a much stronger reduction than Qwen-7B's -26% but far from Qwen-32B's complete collapse. Domain accuracy is perfectly preserved (code: 1.00, math: 1.00, medical: 1.00 at both baseline and step 5000).

The trajectory is oscillatory with a downward trend: 0.799 (base) → 0.759 (500) → 0.941 (1000) → 0.683 (1500) → 0.713 (2000) → 0.598 (2500) → 0.408 (3000) → 0.563 (3500) → 0.365 (4000) → 0.233 (4500) → 0.188 (5000). The temporary increase at step 1000 (EI = 0.941, *above* baseline) and the recovery at step 3500 (0.563 from 0.408) indicate the system is not monotonically approaching a boundary but oscillating within a basin that is progressively shrinking.

The non-degeneracy margin tells the opposite story from Qwen-32B: it *grows* from 3.5×10^{-6} (base) to 4.79 (step 5000), a six-order-of-magnitude increase. The representation is becoming *more* non-degenerate—every SVD direction carries substantial information about every concept—even as EI declines. The EI reduction at 14B is not driven by approach to the non-degeneracy boundary; it reflects a change in the *distribution* of between-class variance across directions (concentration without block-diagonalization).

6.4 Scale pattern

Table 5: Scale ladder within the Qwen-2.5-Coder family. Base EI decreases with scale, and B3 response intensifies monotonically: −26% at 7B, −76% at 14B, −100% at 32B. The transition from partial reduction to complete collapse occurs between 14B and 32B.

Model	Params	Base EI	B3 EI	ΔEI	Final margin
Qwen-7B	7.6B	0.926	0.683	−26%	2.3×10^{-4}
Qwen-14B	14.7B	0.799	0.188	−76%	4.79
Qwen-32B	32.5B	0.667	0.000	−100%	—

6.5 Interpretation: graded susceptibility with a 32B phase boundary

The scale ladder reveals a continuous gradient of susceptibility, not a sharp threshold. The B3 response intensifies monotonically with scale: −26% (7B) → −76% (14B) → −100% (32B). Base EI also decreases monotonically with scale (0.926 → 0.799 → 0.667), consistent with larger Qwen models encoding domains more separately in their activation geometry.

However, the qualitative transition—from partial EI reduction to complete collapse—occurs between 14B and 32B. Qwen-14B's final margin of 4.79 demonstrates that the non-degeneracy condition is firmly satisfied; the representation cannot collapse to EI = 0 without crossing the non-degeneracy boundary, and 5000 steps of B3 training push it *away* from that boundary. Qwen-32B, by contrast, crosses the boundary by step 2500 and never returns.

This supports a two-component interpretation: (1) NL fine-tuning concentrates the between-class variance into fewer directions at all Qwen scales, producing a continuous EI decrease; (2) at 32B, an additional mechanism—plausibly related to the model's larger effective rank or different attention structure—enables the concentration to proceed all the way to block-diagonalization, crossing the non-degeneracy boundary and producing complete collapse. The 14B model approaches but does not cross this boundary, with margin *increasing* throughout training.

7 Pre-Training Composition

If Qwen-32B's pre-training produces code representations already partially separated from natural language, NL fine-tuning might complete the separation, driving block-diagonalization. Models with more deeply integrated code-NL representations would resist disentanglement. We test this by measuring cross-domain activation similarity on base models before any fine-tuning.

For each pair of domains (A, B) in the EI probe set (code, math, medical), we compute four metrics:

1. **Centroid cosine**: $\cos(\bar{\mu}_A, \bar{\mu}_B)$, measuring alignment of domain-mean activations.
2. **Subspace overlap**: fraction of variance in domain A's top-k PCA subspace explained by domain B's top-k subspace ($k = 5$).
3. **Binary LOO-CV**: leave-one-out cross-validated accuracy of a ridge classifier distinguishing domains A and B. Higher accuracy = more separable = less integrated.
4. **Mean cosine**: average pairwise cosine similarity within and between domains.

Table 6: Cross-domain similarity metrics on base models (averages across domain pairs). All models show similar centroid alignment (≈ -0.49) and high binary separability (≥ 0.99). No model stands out as having distinctively separated or integrated domains.

Model	**Centroid cos.**	**Subspace ov.**	**Binary CV**	**EI**
CodeLlama-7B	−0.491	0.248	0.992	0.858
DeepSeek-6.7B	−0.494	0.145	0.992	1.385
Qwen-7B	−0.490	0.259	0.992	0.926
Qwen-14B	−0.495	0.349	1.000	0.799
Qwen-32B	−0.485	0.287	1.000	0.667

7.1 No evidence for Hypothesis B

The cross-domain similarity metrics are remarkably uniform across models. Centroid cosines cluster tightly around -0.49 ($< 2\%$ variation). Binary LOO-CV is at or near ceiling for all models (≥ 0.992). Subspace overlap ranges from 0.145 (DeepSeek) to 0.349 (Qwen-14B), but this variation shows no correlation with either EI or collapse susceptibility: DeepSeek has the lowest overlap (most separated domains) yet the highest EI and no collapse; Qwen-14B has the highest overlap yet does not collapse.

7.2 Limitation: probe domains vs. NL

The EI probe set contains code, mathematics, and medical text—not a dedicated natural language domain. Hypothesis B specifically concerns the separation between code and NL representations. Medical text is a proxy for general NL, but the B3 companion data is diverse NL (not medical text). A direct test would require probes from the B3 NL distribution, computed at the terminal activation layer. The current results rule out gross cross-domain structural differences as predictors but cannot definitively test whether code-NL separation specifically predicts collapse susceptibility.

Within the available data, Hypothesis B finds no support. The models look structurally similar in their domain organization despite dramatically different collapse behavior.

8 Cross-Architecture Replication

To isolate architecture from scale, we apply the B3 protocol to Mistral-7B-v0.3—a model matched in parameter count to CodeLlama and DeepSeek but from a distinct architecture family (Mistral's sliding-window attention, grouped-query attention). Combined with the Paper 44 cross-family results, this provides four independent architecture tests at the ∼7B scale.

8.1 Protocol

Mistral-7B-v0.3 was fine-tuned under the same B3 protocol used throughout: rank-64 QLoRA, all 7 linear target modules, $\alpha = 128$, learning rate 10^{-4}, 5000 steps, code primary + NL secondary companion data. EI diagnostics were recorded every 500 steps. Llama-3.1-8B was excluded due to gated repository access restrictions.

8.2 Results

8.3 Mistral-7B dynamics

The EI trajectory over 5000 steps shows fluctuation without trend: 1.040 (base) → 0.897 (step 500) → 1.059 (1000) → 0.853 (1500) → 1.114 (2000) → 1.141 (2500) → 1.180 (3000) → 0.973 (3500) → 1.056 (4000) → 1.055 (4500) → 1.044 (5000). The standard deviation of the trajectory is 0.098, with no monotonic trend. Domain classification accuracy is perfectly preserved (code: 1.00, math: 0.95, medical: 1.00 at both baseline and step 5000).

The singular value energy *decreases* from 0.073 to 0.023 over training, indicating the adapter is learning (compressing the informative subspace), but this compression does not induce block-diagonalization or EI collapse. This dissociation—adapter learning without

Table 7: Cross-architecture B3 results at ∼7B scale. No model other than Qwen-32B exhibits EI collapse. CodeLlama *increases* in EI; DeepSeek shows a modest decrease; Qwen-7B shows a partial decrease without collapse; Mistral is essentially unchanged. Paper 44 results shown for comparison where new runs were not conducted.

Model	**Family**	**Base EI**	**B3 EI**	**ΔEI**	**Source**
CodeLlama-7B	Llama	0.858	1.09–1.19	+27%–39%	Paper 44
DeepSeek-6.7B	DeepSeek	1.385	1.196	−14%	Paper 44
Qwen-7B	Qwen	0.926	0.683	−26%	This work
Mistral-7B	Mistral	1.040	1.044	+0.4%	This work
Qwen-32B	Qwen	0.667	0.000	−100%	Paper 44

entanglement destruction—is the key observation: the B3 protocol modifies the representation but only the specific combination of Qwen architecture at $\geq$32B scale produces the phase transition.

8.4 Interpretation

Four architecture families at ∼7B scale respond to B3 in qualitatively different ways: increase (CodeLlama), decrease without collapse (DeepSeek, Qwen-7B), no change (Mistral), and complete collapse (Qwen-32B only). The response is not predicted by baseline EI (DeepSeek has the highest at 1.385 yet does not collapse), non-degeneracy margin (Section 5), or pre-training domain structure (Section 7).

The architecture-specificity combined with the scale-specificity (Section 6) converges on a narrow conclusion: the EI collapse under B3 is a property of the Qwen architecture at large scale, not a generic phenomenon of NL fine-tuning. The mechanism is embedded in how Qwen's attention and representation structure organizes cross-domain information at 32B parameters—an architectural fingerprint that is absent or insufficient in other families and at smaller Qwen scales.

9 Reversibility

If the B3 collapse is a phase transition, a natural question is whether it is reversible: can continued fine-tuning with a non-collapsing companion (B2: code+math) recover entanglement?

We attempted to test this by loading the B3-trained QLoRA adapter for Qwen-32B (seed 0, EI = 0.000) and continuing training with B2 data for an additional 5000 steps.

9.1 Adapter failure

The B3 adapter weights from the Paper 44 training runs contain NaN values in every parameter tensor (896 LoRA parameter matrices, all fully NaN). The adapter produces NaN activations on forward pass, preventing any diagnostic computation or further training. This is consistent with the known numerical divergence of QLoRA rank-64 all-linear training on 32B models at learning rate 10^{-4}—the training apparently diverged during the final checkpoint window and saved a corrupted adapter.

9.2 Implications

The reversibility experiment could not be conducted at 32B scale. The question remains open: whether B2 training can recover entanglement from a collapsed representation. A valid test requires either (1) retraining B3 on Qwen-32B with NaN detection and early stopping to produce a non-corrupted collapsed adapter, or (2) testing reversibility on a smaller model that exhibits partial EI reduction under B3. We note that the scale ladder results (Section 6) may identify models with intermediate collapse behavior suitable for this test.

10 Discussion

10.1 The phase transition interpretation

The dynamics data supports interpreting the B3 collapse as a phase transition rather than gradual decay. The evidence:

1. **Sharpness**: B3 and B2 trajectories are indistinguishable through step 1500. The divergence at step 2000 is abrupt (B3/B2 ratio drops from 0.92 to 0.35 in 500 steps).
2. **Irreversibility**: Once permanent collapse occurs, no seed recovers. The system crosses a boundary and does not return.
3. **Transient fluctuations**: Two seeds oscillate near the boundary before permanent collapse, analogous to critical fluctuations near a phase transition.
4. **Condition specificity**: B2 (code+math) never approaches the boundary despite identical training mechanics, indicating the transition requires a specific perturbation direction (NL data).

10.2 Why NL and not math?

The B3/B2 divergence is the strongest constraint on mechanism. Both conditions use the same model, QLoRA configuration, and training infrastructure. The only difference is the companion data: natural language (B3) vs. mathematics (B2). Why does NL drive block-diagonalization while math does not?

One possibility: NL data occupies a representation subspace that overlaps with the cross-domain coupling directions. Fine-tuning on NL pushes the model's representation toward a subspace where code and NL are more separable, which has the side effect of block-diagonalizing the between-class covariance. Math data, being more structurally similar to code (shared formalism, symbolic manipulation), does not push toward this separation.

This interpretation is testable via Experiment 3 (pre-training composition probe): if correct, the base model's code-NL activation similarity should be lower than code-math similarity, indicating that code and NL are already partially separated before fine-tuning.

10.3 Implications for the entanglement theorem

The dynamics data enriches the relationship between the empirical findings and the theoretical framework:

- The theorem proves entanglement is generic (holds for almost every activation matrix). The B3 collapse shows that fine-tuning can move the model off the generic set—but only in specific architecture-scale combinations. Paradoxically, the model with the *largest* non-degeneracy margin is the one that collapses (Section 5), suggesting the mechanism is not a small perturbation crossing a nearby boundary but a qualitative reorganization of the representation geometry.
- The temporary recoveries provide the first empirical evidence of oscillation near the non-degeneracy boundary, supporting the geometric interpretation of Assumption 1 as a genuine boundary rather than a mathematical technicality.
- The preserved V-purity during collapse confirms the discrimination–activation dissociation at the dynamical level, strengthening the theorem's distinction between how classifiers *use* directions and what information those directions *carry*.
- The cross-architecture results (Section 8) show four qualitatively different responses to B3 at $\sim$7B scale—increase, decrease, no change, and (only at 32B) collapse—indicating that the non-degeneracy condition's violation is architecture-specific, not a generic consequence of NL fine-tuning.

10.4 Limitations

The dynamics analysis has three limitations. First, the checkpoint diagnostics do not include per-cell damage matrices, so we cannot determine the order in which specific domain pairs lose their cross-coupling. This would require reloading intermediate adapter checkpoints and re-running the full damage matrix computation (a GPU-dependent analysis). Second, the checkpoint granularity is 500 steps, which may be too coarse to resolve the exact transition dynamics. Third, the effective rank of 5 (vs. theoretical 7) in the base model suggests that Qwen-32B's representation geometry is already atypical at the measurement layer, and the collapse dynamics may not generalize to models with full-rank informative subspaces.

11 Future Work

1. **Per-cell collapse ordering**: Load intermediate checkpoint adapters and compute the full damage matrix at each step to determine whether specific domain pairs collapse before others.
2. **Reversibility at smaller scale**: Test whether B2 training can recover entanglement from a partially reduced model (e.g., Qwen-14B after B3, EI = 0.188). The 32B reversibility test was blocked by adapter corruption (Section 9), but 14B's strong reduction without collapse makes it an ideal candidate.
3. **Qwen-20B or Qwen-24B**: The scale ladder reveals a continuous gradient with a qualitative phase boundary between 14B and 32B. An intermediate scale point would narrow the boundary location.
4. **NL-specific composition probes**: The current cross-domain similarity analysis (Section 7) uses code/math/medical probes. Direct probes from the B3 NL distribution would provide a sharper test of Hypothesis B.
5. **Llama-3.1-8B**: Excluded from the cross-architecture study due to gated repository access. This model would add a fourth family at matched scale.
6. **Finer checkpoint granularity**: The current 500-step resolution may be too coarse to resolve the exact transition dynamics. Recording diagnostics every 100 steps during the critical 1500–3500 window would better characterize the phase transition.

12 Conclusion

Natural language fine-tuning produces a sharp phase transition in the activation geometry of Qwen-2.5-Coder-32B, driving entanglement intensity from 0.622 to 0.000 within 2000–3500 training steps while leaving the discrimination geometry unchanged. The collapse is condition-specific (B2 code+math never collapses), architecture-specific (absent in CodeLlama, DeepSeek, and Mistral at 7B), and irreversible once permanent. Temporary recoveries in two of eight seeds provide evidence of oscillation near the non-degeneracy boundary predicted by the entanglement theorem.

Of the three candidate mechanisms, two are ruled out by the data. Hypothesis A (non-degeneracy margin) is refuted: Qwen-32B has the *largest* margin of all tested models, not the smallest. Hypothesis B (pre-training composition) finds no support: all models show similar cross-domain activation structure. The remaining explanation, Hypothesis C (scale dependence), is confirmed by the complete scale ladder: base EI decreases with Qwen model size ($0.926 \rightarrow 0.799 \rightarrow 0.667$), and the B3 response intensifies continuously from -26% at 7B through -76% at 14B to -100% at 32B. The transition is graded, not threshold-like, but a qualitative phase boundary exists between 14B and 32B: at 14B the non-degeneracy margin *grows* during training (from 10^{-6} to 4.79), preventing complete collapse, while at 32B the margin is crossed and the collapse is permanent.

The cross-architecture results establish that the effect is narrow: four architecture families at ∼7B scale produce four qualitatively different responses to B3 (increase, decrease, no change, and—only at 32B Qwen—collapse). Structural entanglement, while geometrically generic, is dynamically robust in most architectures. Its destruction requires a specific combination of architecture, scale, and fine-tuning direction that converges on Qwen at large scale.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proc. ACL*, 2021.

Jeremy McEntire. Structural entanglement in the informative subspace: Eight experiments on why every direction carries every concept. Zenodo, 2026. doi:10.5281/zenodo.18880969.

Jeremy McEntire. The entanglement theorem: Structural concept coupling as a geometric consequence of high-dimensional encoding. Zenodo, 2026. doi:10.5281/zenodo.18880971.

Jeremy McEntire. Entanglement-optimal fine-tuning: Crosstalk-guided companion selection and complement-subspace regularization for code models. Working paper, 2026.

Aaron Mueller, Andrew Lee, Shruti Joshi, Ekdeep Singh Lubana, Dhanya Sridhar, and Patrik Reizinger. From isolation to entanglement: When do interpretability methods identify and disentangle known concepts? *arXiv preprint arXiv:2512.15134*, 2025.

Ruiqi Zhong, Tao Lei, Diyi Yang, and Jacob Steinhardt. Do fine-tuning and retrieval change the activation patterns of pre-trained language models? *arXiv preprint arXiv:2510.09359*, 2025.