Entanglement Under Fine-Tuning:

Scale-Dependent Disentanglement, Crosstalk-Guided Companion Selection, and the Creative Writing Falsification

Jeremy McEntire*

Abstract

Fine-tuning changes what a model knows. It also changes how that knowledge is organized—and the reorganization follows a scale-dependent pattern that no existing theory predicts.

We study what happens to structural entanglement—the property that every informative direction in a transformer's activation space carries all concept dimensions simultaneously in its pre-trained state—when models are fine-tuned with QLoRA (Quantized Low-Rank Adaptation—a method for fine-tuning large models using reduced-precision weights with low-rank update matrices) across six conditions. The experiments span three scales (Qwen-2.5-Coder at 7B, 14B, and 32B), four architecture families (Qwen, Mistral, DeepSeek, CodeLlama), and include a lambda sensitivity sweep for complement-subspace regularization (CSR—a penalty that discourages the model from changing its representations in directions unrelated to the target domain) and a reversibility test.

Four findings. First, disentanglement effectiveness is monotonically scale-dependent within the Qwen family: entanglement intensity (EI) after fine-tuning is 0.641 ± 0.099 at 7B (32% reduction from baseline, 3 seeds), 0.063 ± 0.109 at 14B (93% reduction, 3 seeds), and 0.000 ± 0.000 at 32B (100% reduction, 8 seeds). The transition from partial to complete disentanglement lies between 7B and 14B. Second, crosstalk structure partially predicts fine-tuning interference—math companions retain 11.6 percentage points more coding performance than natural language companions ($p = 0.016$)—but the creative writing condition (lowest measured crosstalk, CT = 0.0015) produces the *worst* coding performance among non-curriculum conditions, honestly falsifying naive crosstalk minimization. Third, CSR is robust to hyperparameter choice: final EI varies by less than 0.01 across a 10× lambda range (0.03–0.3). Fourth, preliminary evidence ($N = 1$) suggests disentanglement may be irreversible at this scale: sequential B3→B2

*Correspondence: jmc@cageandmirror.com

fine-tuning on Qwen-14B drives EI from 0.279 to 0.000 within 1000 steps rather than recovering entanglement.

At the 7B scale, cross-architecture results are mixed: Qwen shows 32% EI reduction, DeepSeek 24%, Mistral 10%, and CodeLlama shows 31% *anti*-disentanglement. Disentanglement at 7B is architecture-dependent; at 14B+ within Qwen, it is robust and deterministic.

1 Introduction

Fine-tuning changes what a model knows. It also changes how that knowledge is organized—and the reorganization follows a scale-dependent pattern that no existing theory predicts.

The parent paper in this series [McEntire, 2026] established *structural entanglement*: when k concepts are encoded in $d \gg k$ dimensions via multi-concept ridge regression, every informative direction carries all k concepts simultaneously. The entanglement intensity is characterized by EI $\approx (r - d_{\max})/d_{\max}$, where r is the informative rank and $d_{\max}$ is the largest concept subspace dimension. In plain terms, the Entanglement Index measures how much removing one concept's direction damages other concepts. Values above 1.0 mean the direction hurts other concepts more than its own; 0.0 means complete disentanglement. This property holds under a generic non-degeneracy condition satisfied by Lebesgue-almost-every activation matrix, and was confirmed across four transformer architectures spanning a 60× parameter range.

The present work asks what happens to entanglement when you fine-tune.

Understanding how fine-tuning changes entanglement has direct practical consequences. If a safety team removes a dangerous capability by erasing its direction, entanglement determines whether that erasure damages unrelated capabilities. If a model editor modifies factual knowledge, entanglement determines the collateral damage. The structure of entanglement after fine-tuning governs what is safe to edit and what is not.

The question matters for three reasons. First, practitioners routinely fine-tune large models on domain-specific data—but the effect on the model's internal representational organization is unmeasured and unpredicted. If fine-tuning reorganizes concept geometry, current interpretability methods calibrated on base models may not apply to fine-tuned variants. Second, the *choice of companion data*—what else you train on besides the target domain—is largely heuristic. Practitioners know that adding math helps coding models and that irrelevant data hurts, but the mechanism is unclear. Third, the entanglement theorem offers a framework for predicting interference: the crosstalk structure of the activation space should predict which companion domains will help and which will hurt.

We test these ideas across three scales (7B, 14B, 32B), four architecture families (Qwen, Mistral, DeepSeek, CodeLlama), and six fine-tuning conditions. The headline finding is a scale-dependent disentanglement pattern: Qwen-32B reaches EI = 0.000 across all 8 seeds, Qwen-14B reaches 0.063 ± 0.109, and Qwen-7B reaches only 0.641 ± 0.099. The transition between partial and complete disentanglement lies between 7B and 14B parameters. No existing theory predicts this gradient.

The companion-selection results partially validate the crosstalk prediction: math companions outperform natural language companions by 11.6 percentage points on HumanEval+ ($p = 0.016$). But the creative writing condition—the companion with the *lowest* measured crosstalk (CT = 0.0015, block-diagonal ratio 0.99)—produces the worst performance among non-curriculum conditions. This honestly falsifies naive crosstalk minimization and constrains the framework: low interference is necessary but not sufficient. A companion must also provide transferable inductive bias.

Two additional results complete the picture. CSR—a geometric regularization that penalizes activation drift in non-target directions—is robust to a 10× lambda sweep, eliminating hyperparameter sensitivity as a practical concern. And a reversibility test on Qwen-14B shows that disentanglement deepens rather than reverses under continued fine-tuning with a different companion, suggesting a one-way ratchet in representational geometry.

2 Background

2.1 Structural entanglement

Structural entanglement was discovered, formalized, and replicated in McEntire [2026]. The key objects:

- The **weight matrix** $W \in \mathbb{R}^{d \times c}$ from multi-output ridge regression over factorial probes spanning k concept dimensions with c total classes.
- The **SVD** $W = U\Sigma V^\top$, where columns of U are activation-space directions and rows of V are concept loadings.
- The **removal damage matrix** $D \in \mathbb{R}^{r \times k}$: entry D_{ij} is the accuracy drop on concept j when direction i is removed.
- **Entanglement intensity** $\text{EI} = \sum_i \sum_{j \neq \text{own}(i)} D_{ij} / \sum_i D_{i,\text{own}(i)}$, the ratio of off-diagonal to diagonal damage.

The entanglement theorem proves that under a generic non-degeneracy condition, $\text{EI} \approx (r - d_{\max})/d_{\max}$, which exceeds 1 whenever $d_{\max} < r/2$. The specialist bound (Corollary 3) states that EI scales superlinearly with k: distributing concepts across specialist modules with $k_i \ll k_{\text{total}}$ reduces entanglement. We refer the reader to McEntire [2026] for derivations

and proofs; the present paper is purely empirical.

2.2 QLoRA fine-tuning

All experiments use QLoRA [Dettmers et al., 2023]: 4-bit NF4 quantization with double quantization, LoRA rank 64, $\alpha = 128$, targeting all 7 linear modules (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj), dropout 0.05, cosine learning rate schedule with 3% warmup, batch size 1 with gradient accumulation 16, gradient checkpointing, learning rate 1×10^{-4}, max sequence length 2048, and 5000 training steps.

Training data. Code: `code_search_net` (Python, JavaScript) + `CodeAlpaca-20k`. Math: `gsm8k` + `competition_math`. Natural language: `wikipedia` (en, 2023-11-01) + `cnn_dailymail`. Creative writing: fiction and narrative text.

2.3 The six conditions

Table 1 defines the six fine-tuning conditions, each designed to test a specific prediction from the crosstalk framework.

Table 1: Fine-tuning conditions. All use rank-64 QLoRA, all linear layers, 5000 steps ("strong intervention").

Condition	Training Data	Rationale
B0: Baseline	None (base model)	Published reference
B1: Code-only	Code data	Single-domain control
B2: Code+Math	70% code, 30% math	Low-crosstalk companion (CT = 0.088)
B3: Code+NL	70% code, 30% NL	High-crosstalk companion
B4: Code+Math+CSR	Same as B2 + CSR reg.	Complement-subspace regularization
B5: Curriculum	1000 steps math, then code	Sequential domain transfer
B6: Code+Creative	70% code, 30% creative	Lowest-crosstalk companion (CT = 0.0015)

3 Methods

3.1 Models

We test six models across three architecture families and three scales:

- **Qwen family (scale ladder):** Qwen-2.5-Coder-7B (7.6B params, hidden dim 3584), Qwen-2.5-Coder-14B (14.7B params), Qwen-2.5-Coder-32B (32.5B params, hidden dim 5120).
- **Cross-architecture at ∼7B:** Mistral-7B-v0.3 (7.2B params), DeepSeek-Coder-6.7B (6.7B params), CodeLlama-7B (6.7B params).

All Qwen models use the Qwen-2.5-Coder variant. Cross-architecture models are tested at B3 only (code + NL companion), matching the condition that produces the strongest disentanglement in Qwen. All models are open-weight and publicly available on HuggingFace at the time of writing.

3.2 Measurement protocol

Entanglement intensity. For each model and condition, we construct factorial probes varying domain (code, math, medical; 3 levels), register (formal, informal, technical, conversational; 4 levels), and reasoning shape (hierarchical, causal, constraint, evidence; 4 levels), yielding $3 \times 4 \times 4 = 48$ probes. Activations are captured at the terminal transformer layer. We fit a multi-output ridge classifier ($\alpha = 1.0$) with leave-one-out cross-validation and per-fold centering, compute the full SVD, and derive the removal damage matrix. EI is computed as the ratio of off-diagonal to diagonal total damage. Measurements are taken before fine-tuning and at 500-step intervals during training.

V-matrix purity. The V-matrix from the SVD captures discrimination geometry—how the classifier assigns each direction to concepts. V-purity (the fraction of each direction's classifier weight attributable to a single concept) quantifies the degree to which each direction loads on a single concept. It is reported as a stability check: if V-purity changes dramatically during training, the probe infrastructure may no longer be measuring the same quantities.

Domain accuracy. Leave-one-out cross-validated accuracy on the domain classification task. Reported as a fidelity check: disentanglement should not destroy the model's ability to distinguish domains.

Coding performance. Evaluated on HumanEval [Chen et al., 2021] (164 problems) and HumanEval+ [Liu et al., 2024] (164 problems, stricter tests) using the EvalPlus framework. Raw completion mode (no chat template), greedy decoding ($T = 0$), 768 max new tokens. All evaluations are deterministic; the B0 baseline reproduces the published 65.9% HumanEval pass@1.

3.3 Seeds and replication

- **Qwen-32B:** 8 seeds (0–7) for B3; seed 0 for B1, B2, B4, B5, B6; seed 0 for CSR lambda sweep (3 lambda values).
- **Qwen-14B:** 3 seeds (42, 0, 1) for B3; 1 seed for reversibility.
- **Qwen-7B:** 3 seeds (42, 0, 1) for B3.
- **Cross-architecture:** 3 seeds for Mistral-7B (42, 0, 1); 2 seeds for DeepSeek-6.7B (0, 1); 2 seeds for CodeLlama-7B (0, 1).

3.4 CSR lambda sweep

Complement-subspace regularization (Section 6) adds a loss term penalizing activation drift in non-target directions. The single free parameter is λ, the regularization strength. We sweep $\lambda \in \{0.03, 0.1, 0.3\}$ on Qwen-32B, seed 0, under the B4 condition.

3.5 Reversibility protocol

To test whether disentanglement is reversible, we take a B3-trained adapter (Qwen-14B, seed 0, post-B3 EI = 0.279) and continue training with B2 data (code + math) for an additional 1500 steps, recording EI at 500-step intervals.

3.6 Block-diagonal structure measurement

For the crosstalk analysis, we expand the factorial design to six domains: code, math, formal logic, medical, legal, and creative writing ($6 \times 4 \times 4 = 96$ probes per domain, 480 total). The pairwise crosstalk between code and each companion domain is computed from the 6×6 removal damage matrix.

Definition 1 (Symmetric crosstalk)**.** *For domains A and B with normalized damage values:*

$$CT(A, B) = \tfrac{1}{2}(D_{A\to B} + D_{B\to A}).$$

Definition 2 (Block-diagonal ratio)**.**

$$BDR(A, B) = \frac{D_{A\to A} + D_{B\to B}}{D_{A\to A} + D_{A\to B} + D_{B\to A} + D_{B\to B}}.$$

A ratio of 1.0 means perfectly block-diagonal; 0.5 means off-diagonal damage equals diagonal damage.

4 Scale-Dependent Disentanglement

This is the headline result. Within the Qwen-2.5-Coder family, disentanglement effectiveness under B3 fine-tuning increases monotonically with model scale, following a pattern that transitions from partial at 7B to complete and deterministic at 32B.

4.1 The scale ladder

Table 2 reports the aggregate results. All experiments use the B3 condition (70% code, 30% NL).

Table 2: Scale ladder within the Qwen-2.5-Coder family under B3 fine-tuning. EI decreases monotonically with scale. The transition from partial to complete disentanglement occurs between 7B and 14B.

Model	Seeds	EI (before)	EI (after)	V-purity (after)	Domain acc
Qwen-7B	3	0.945 ± 0.017	0.641 ± 0.099	0.643 ± 0.020	0.950
Qwen-14B	3	0.810 ± 0.010	0.063 ± 0.109	0.620 ± 0.039	0.983
Qwen-32B	8	0.622 ± 0.000	0.000 ± 0.000	0.622 ± 0.022	0.958

Note: Lower EI indicates more disentangled representations: 0.0 is complete disentanglement, values above 1.0 indicate high entanglement.

The numbers are unambiguous:

- **7B:** EI drops from $\sim$0.95 to 0.641 ± 0.099, a 32% reduction. Disentanglement is partial. The model's representational space is small enough that code and NL share too many directions for full separation.
- **14B:** EI drops to 0.063 ± 0.109, a 93% reduction. Two of three seeds reach EI = 0.000; one seed (42) reaches 0.188. The additional capacity allows the model to separate code and NL into nearly orthogonal subspaces.
- **32B:** EI reaches exactly 0.000 across all 8 seeds with zero variance. Complete disentanglement is deterministic at this scale.

The complete collapse of entanglement intensity to zero across all eight seeds at 32B demonstrates that entanglement is not geometrically forced. Models with sufficient capacity and appropriate training signal can represent concepts independently. The concentration-of-measure bound from the entanglement theorem [McEntire, 2026] remains correct—it bounds EI for a given informative subspace rank r—but the model determines r. At 32B scale with NL companion training, the model expands its informative subspace sufficiently to disentangle. This reframes entanglement from a fundamental geometric limit to a diagnostic of the model's compression strategy: high EI indicates aggressive compression into a low-rank

subspace; EI collapse indicates the model has allocated sufficient representational capacity to separate concepts.

V-purity remains stable across scales (∼0.62), indicating the discrimination geometry—how the classifier assigns directions to concepts—does not change. Only the activation geometry changes: the degree to which each direction carries information about concepts it is not assigned to.

4.2 Per-seed results

Table 3 reports every individual seed to demonstrate reproducibility and characterize variance.

Table 3: Per-seed B3 results across the Qwen scale ladder. All seeds show the same qualitative pattern within each scale. At 32B, the result is deterministic: zero EI, zero variance.

Model	**Seed**	**EI (before)**	**EI (after)**	**V-purity (after)**	**Domain acc**
Qwen-7B	42	0.926	0.683	0.628	0.950
Qwen-7B	0	0.955	0.712	0.635	0.950
Qwen-7B	1	0.955	0.529	0.666	0.950
Qwen-14B	42	0.799	0.188	0.648	1.000
Qwen-14B	0	0.816	0.000	0.638	0.967
Qwen-14B	1	0.816	0.000	0.575	0.983
Qwen-32B	0	0.622	0.000	0.625	0.917
Qwen-32B	1	0.622	0.000	0.636	0.983
Qwen-32B	2	0.622	0.000	0.648	0.950
Qwen-32B	3	0.622	0.000	0.627	0.967
Qwen-32B	4	0.622	0.000	0.601	0.950
Qwen-32B	5	0.622	0.000	0.598	0.933
Qwen-32B	6	0.622	0.000	0.649	0.967
Qwen-32B	7	0.622	0.000	0.595	1.000

Several features are worth noting. The 7B seeds show moderate variance (0.529–0.712), with seed 1 producing the deepest disentanglement. The 14B seeds split: seed 42 stops at 0.188 while seeds 0 and 1 reach zero. The 32B seeds are identical in their terminal EI—all exactly zero—despite different V-purity values (0.595–0.649) and domain accuracies (0.917–1.000). The representational reorganization is the same; only the details of how the classifier uses the reorganized space vary.

4.3 Training dynamics

The temporal structure reveals that disentanglement is not gradual decay but a process with characteristic timing that depends on scale.

At 7B (seed 1, which shows the clearest descent), EI starts high (∼1.1), holds through step 2500, then drops steadily to 0.529 by step 5000. Disentanglement is a late-training phenomenon at this scale—the first half of training increases entanglement before the model learns to separate domains.

At 14B (seeds 0 and 1), EI reaches 0.000 by step 3000–3500 and stays there for the remaining 1500–2000 steps. The trajectory is oscillatory with a downward trend, including a temporary increase above baseline at step 1000 (EI = 0.941). The system oscillates within a basin that is progressively shrinking.

At 32B (seed 0), EI reaches 0.000 by step 2500. The larger model disentangles fastest. The B3 vs. B2 trajectory comparison (Table 4) shows that the two conditions track together through step 1500, then diverge sharply: B3 collapses to zero by step 3500 while B2 fluctuates around 0.40 indefinitely. Same model, same infrastructure, different companion data.

Table 4: Mean EI trajectory (± 1 SD) across 8 seeds for B3 and B2 conditions on Qwen-32B. The conditions are indistinguishable through step 1500, then diverge at step 2000. B2 never approaches zero.

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

Two of eight 32B seeds (25%) exhibit temporary recovery after initial collapse: seed 5 reaches EI = 0.000 at step 2000, recovers to 0.200 at step 2500, then collapses permanently at step 3000; seed 6 reaches 0.000 at step 2500, recovers to 0.452 at step 3000, then collapses permanently at step 3500. These transient oscillations are consistent with the system fluctuating near the non-degeneracy boundary predicted by the entanglement theorem.

4.4 Cross-architecture comparison at 7B

Is disentanglement a generic consequence of NL fine-tuning, or is it architecture-specific? Table 5 reports the B3 response across four 7B-class architectures.

Table 5: Cross-architecture B3 results at ∼7B scale. No model other than Qwen-32B exhibits complete EI collapse. CodeLlama shows *anti*-disentanglement. Only Qwen shows meaningful disentanglement at this scale.

Architecture	**Seeds**	**EI (after)**	**EI (before)**	**ΔEI**	**Response**
Qwen-7B	3	0.641 ± 0.099	0.945	−32%	Partial reduction
DeepSeek-6.7B	2	1.047 ± 0.210	1.374	−24%	Moderate reduction
Mistral-7B	3	0.944 ± 0.098	1.044	−10%	Minimal change
CodeLlama-7B	2	1.143 ± 0.070	0.874	+31%	Anti-disentanglement

Table 6: Per-seed cross-architecture B3 results. CodeLlama consistently shows EI *increase* across both seeds.

Architecture	**Seed**	**EI (before)**	**EI (after)**	**V-purity (after)**	**Domain acc**
Qwen-7B	42	0.926	0.683	0.628	0.950
Qwen-7B	0	0.955	0.712	0.635	0.950
Qwen-7B	1	0.955	0.529	0.666	0.950
Mistral-7B	42	1.040	1.044	0.633	0.983
Mistral-7B	0	1.046	0.941	0.639	1.000
Mistral-7B	1	1.046	0.848	0.632	1.000
DeepSeek-6.7B	0	1.376	1.196	0.637	1.000
DeepSeek-6.7B	1	1.372	0.899	0.631	1.000
CodeLlama-7B	0	0.874	1.093	0.625	0.983
CodeLlama-7B	1	0.874	1.192	0.623	0.983

The four architecture families at ∼7B produce four qualitatively different responses to the same B3 protocol: meaningful reduction (Qwen), moderate reduction with high variance (DeepSeek), minimal change (Mistral), and outright increase (CodeLlama). The response is not predicted by baseline EI (DeepSeek has the highest at 1.374 yet does not collapse), nor by the non-degeneracy margin (all ≤14B models have margins within a few ULPs of zero).

The interpretation is straightforward: 7B is below the threshold for architecture-independent disentanglement. At this scale, the fine-tuning response depends on architecture-specific details—how attention is organized, how cross-domain information is represented, what the pre-training mixture looks like. At 14B+ within the Qwen family, sufficient capacity enables robust disentanglement regardless of random seed.

The CodeLlama result directly challenges any claim of universality at the 7B scale. Rather than dismissing it as noise, we note that CodeLlama's architecture and training data distribution differ substantially from the Qwen family. Whether this anti-disentanglement reflects a genuine architectural difference or a measurement artifact at the boundary of the scale threshold remains an open question that future work must address.

4.5 Nondegeneracy margins

Table 7 reports the pre-training nondegeneracy margins for each base model. The margin quantifies how far a model sits from the boundary of the entanglement theorem's non-degeneracy condition.

Table 7: Pre-training nondegeneracy margins. Qwen-32B has a dramatically higher margin than all other models, yet is the only model that completely disentangles—refuting the hypothesis that proximity to the non-degeneracy boundary predicts collapse susceptibility.

Model	EI (base)	Nondeg margin	Weakest category	V-purity
Qwen-7B	0.926	0.000193	shape	0.615
Qwen-14B	0.799	0.000004	shape	0.616
Qwen-32B	0.667	1.781	register	0.635
DeepSeek-6.7B	1.386	0.000021	register	0.623
CodeLlama-7B	0.858	0.000000	register	0.627

Qwen-32B's margin (1.781) is four orders of magnitude larger than any other model. All $\leq$14B models have near-zero margins, yet none collapses under B3. The direction of the effect is the opposite of what proximity-to-boundary theories would predict: the model with the *strongest* non-degeneracy is the one whose entanglement is destroyed. 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.

4.6 Interpretation

The scale dependence is the central finding. Three observations constrain the mechanism:

1. **The gradient is continuous.** The B3 response intensifies monotonically: -32% (7B), -93% (14B), -100% (32B). There is no sharp threshold—the transition is graded.
2. **Base EI decreases with scale.** Before any fine-tuning, EI is 0.926 (7B), 0.799 (14B), 0.622 (32B). Larger Qwen models encode domains more separately in their base activation geometry.
3. **The 32B phase transition is qualitatively different.** At 14B, the non-degeneracy margin *grows* during training (from 3.5×10^{-6} to 4.79), preventing complete collapse. At 32B, the margin is crossed and the collapse is permanent. An additional mechanism—plausibly related to the model's larger effective rank or different attention structure—enables the concentration to proceed to block-diagonalization.

No existing theory predicts this scale dependence. The entanglement theorem characterizes EI for a *fixed* encoding; it does not predict how fine-tuning changes EI as a function of model size.

5 Crosstalk-Guided Companion Selection

The crosstalk matrix—derived from the removal damage matrix on Qwen-2.5-Coder-7B across six knowledge domains—provides a per-domain-pair measure of activation-space interference. Table 8 reports the pairwise crosstalk between code and five companion domains.

Table 8: Pairwise crosstalk between code and five companion domains in Qwen-2.5-Coder-7B. Domains ranked by symmetric crosstalk coefficient. Lower crosstalk = more block-diagonal = less predicted interference.

Companion Domain	**Sym. Crosstalk**	**Block-Diag. Ratio**	**Asymmetry**
Creative writing	0.0015	0.991	1.00
Math	0.088	0.750	1.00
Formal logic	0.092	0.747	1.00
Medical	0.170	0.487	1.00
Legal	0.296	0.499	0.23

The crosstalk structure spans two orders of magnitude: code–creative-writing is nearly perfectly block-diagonal (CT = 0.0015, BDR = 0.991), while code–legal shows heavy bidirectional interference (CT = 0.296, BDR = 0.499). Math and formal logic cluster at intermediate values (CT $\approx$ 0.09) with perfectly unidirectional asymmetry: removing math-informative directions damages code, but removing code-informative directions does not damage math.

5.1 Coding performance under varied companions

Table 9 reports HumanEval and HumanEval+ results under strong intervention on Qwen-32B.

Table 9: HumanEval results (pass@1) under strong intervention on Qwen-2.5-Coder-32B. Rank 64, all linear layers, 5000 steps. Evaluation is deterministic (greedy decoding, $T = 0$). Raw counts on 164 HumanEval+ problems shown alongside percentages.

Condition	**HE**	**HE+**	**HE+ (raw)**	**Δ HE+**	**95% CI (Δ)**
B0: Baseline	65.9%	58.5%	96/164	—	—
B4: Code+Math+CSR	57.9%	49.4%	81/164	−9.1	[−16.7, −1.5]
B2: Code+Math	56.7%	48.2%	79/164	−10.3	[−18.0, −2.6]
B1: Code-only	54.9%	47.6%	78/164	−10.9	[−18.6, −3.2]
B6: Code+Creative	45.1%	41.5%	68/164	−17.0	[−27.7, −6.3]
B3: Code+NL	42.7%	36.6%	60/164	−21.9	[−29.1, −14.7]
B5: Curriculum	12.8%	10.4%	17/164	−48.1	[−53.6, −42.6]

Note: Raw counts (correct/total) are reported alongside pass rates.

The companion-selection results partially validate the crosstalk prediction:

1. **B2 > B3** (HumanEval+ 48.2% vs. 36.6%, $z = 2.15$, $p = 0.016$, one-tailed). The math companion (CT = 0.088) retains 11.6 percentage points more coding performance than the NL companion. B3 loses nearly twice as much performance as B2 relative to baseline. This is statistically significant and practically large.
2. **B4 is the best fine-tuned condition** (49.4%, 81/164). CSR preserves coding performance better than unrestricted fine-tuning.
3. **B5 is catastrophic** (10.4%, 17/164). Sequential math-then-code training destroys coding ability. The 1000 math-only steps move the model to a region of weight space that 4000 subsequent code steps cannot recover.

5.2 The creative writing falsification

The B6 result deserves its own subsection because it is a first-class finding, not a footnote.

Code–creative-writing has the lowest crosstalk of any measured pair: CT = 0.0015, BDR = 0.991. A naive reading of the crosstalk-minimization framework predicts this should be the *best* companion—or at minimum, a harmless one. It is neither. B6 achieves HumanEval+ 41.5% (68/164), a 17.0 percentage point drop from baseline ($z = -3.08$, $p = 0.001$). It is the second-worst condition, beating only NL (B3) and curriculum (B5).

This result honestly falsifies naive crosstalk minimization. The companion with the lowest interference produces among the worst outcomes.

The falsification constrains the framework in a useful way. The crosstalk matrix measures *interference*—how much removing one domain's directions damages another domain's classification. Low interference means the companion will not overwrite the target domain's representations. But a fine-tuning companion must also provide *transferable inductive bias*—reasoning structure that helps the target domain. Math shares formal reasoning with code (CT = 0.088, HumanEval+ 48.2%). Creative writing shares no such structure (CT = 0.0015, HumanEval+ 41.5%). The 30% of training steps spent on creative writing data are not merely wasted—they actively displace coding representations without providing compensating structure.

The singular value energy confirms this: B6's SV energy (0.041) is moderate—between B1 (0.014) and B5 (0.646)—indicating focused but unhelpful weight changes. The NL companion causes 1000× more weight displacement than CSR (1.096 vs. 0.001), consistent with NL's broad activation footprint overwriting code-specific representations.

Table 10: Singular value energy of weight changes under strong intervention. SV energy is strongly anticorrelated with coding performance (Spearman $r_s = -0.94$). B6 reinforces the pattern: focused weight changes are necessary but not sufficient.

Condition	SV Energy	HE+	Interpretation
B4: Code+Math+CSR	0.001	49.4%	CSR constrains changes maximally
B2: Code+Math	0.006	48.2%	Math companion keeps changes focused
B1: Code-only	0.014	47.6%	Code-only spreads more
B6: Code+Creative	0.041	41.5%	Low crosstalk, no transferable structure
B5: Curriculum	0.646	10.4%	Sequential: massive displacement
B3: Code+NL	1.096	36.6%	NL spreads changes broadly

5.3 Statistical methodology

All HumanEval evaluations use greedy decoding ($T = 0$), producing deterministic outcomes. We treat the 164 HumanEval+ problems as a finite sample and use the two-proportion z-test:

For the B2 vs. B3 gap (79/164 vs. 60/164):

$$z = \frac{0.482 - 0.366}{\sqrt{\hat{p}(1-\hat{p})(2/164)}} = 2.15, \quad p = 0.016 \text{ (one-tailed)}$$

where $\hat{p} = (79 + 60)/(2 \times 164) = 0.424$ is the pooled proportion.

The B6 vs. B0 gap (68/164 vs. 96/164): $z = -3.08$, $p = 0.001$.

Caveat. We note that HumanEval is a fixed benchmark, not a random sample, and two-proportion z-tests assume independent Bernoulli draws. We report these tests as conventional practice but emphasize the effect sizes (raw count differences) as the primary evidence. Bootstrap confidence intervals would be more appropriate; we leave this refinement to future work. The test quantifies how surprising the observed difference would be under a null model of equal capability; it does not establish that the difference generalizes to arbitrary coding tasks.

6 CSR Regularization

6.1 Mechanism

Standard fine-tuning updates weights in all directions, including those that serve non-target domains. Complement-subspace regularization (CSR) constrains learning to target-domain directions while penalizing drift in the complement subspace.

1. Before training, capture activation vectors $\{h_i^{(0)}\}$ from the base model on the full set of factorial probes ($N = 480$ probes, hidden dimension $d = 5120$ for 32B).
2. Compute the target-domain subspace $S_{\text{target}} \in \mathbb{R}^{d \times p}$ via truncated SVD of the target-domain activation matrix, retaining the top p singular vectors capturing 90% of the variance. In our experiments, $p = 47$ for the code domain, so the complement has dimension $d - p = 5073$.
3. Compute the complement projector $P_\perp = I - S_{\text{target}} S_{\text{target}}^\top$.
4. Every 50 training steps, run the current model on the factorial probes to obtain $\{h_i^{(t)}\}$ and add:

Intuitively, CSR works by penalizing changes to the model's internal representations in directions that do not correspond to the target domain. The penalty is proportional to how much the fine-tuned model's activations drift from the base model in the complement of the target subspace:

$$\mathcal{L}_{\text{CSR}} = \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} \| P_\perp (h_i^{(t)} - h_i^{(0)}) \|^2 \tag{1}$$

CSR operates in *activation space* rather than weight space, targeting the geometric structure that the entanglement theorem describes. EWC [Kirkpatrick et al., 2017] asks "which weights matter?"; CSR asks "which activation directions are mine?"

6.2 Lambda sensitivity

Table 11 reports the CSR lambda sweep on Qwen-32B, seed 0, under the B4 condition.

Table 11: CSR lambda sweep on Qwen-32B, B4 condition, seed 0. Final EI varies by less than 0.01 across a 10× lambda range. CSR is robust to hyperparameter choice.

Lambda	EI (before)	EI (after)	V-purity (after)	Domain acc	Final SR loss
0.03	0.621	0.105	0.682	1.000	7.221
0.10	0.255	0.115	0.650	1.000	7.221
0.30	0.621	0.105	0.682	1.000	72.210

The result is striking: across a 10× lambda range (0.03 to 0.3), final EI converges to the same value (0.105–0.115, a difference of 0.01). Domain accuracy is preserved at 1.000 across all lambdas. V-purity is slightly higher at lambda = 0.03 and 0.30 (0.682 vs. 0.650 at lambda = 0.10).

The lambda = 0.03 and lambda = 0.30 runs produce identical EI trajectories step-by-step (Table 12), differing only in the magnitude of the SR loss term (which scales linearly with lambda: the 0.30 run has SR losses exactly 10× the 0.03 run). This is consistent with the

CSR loss being a small perturbation on the overall training loss at these scales, affecting the loss magnitude but not the gradient direction.

Table 12: EI trajectory during CSR training at three lambda values. The lambda = 0.03 and 0.30 trajectories are identical; lambda = 0.10 follows a different path but converges to the same endpoint.

Step	$\lambda = 0.03$	$\lambda = 0.10$	$\lambda = 0.30$
500	0.220	0.064	0.220
1000	0.582	0.196	0.582
1500	0.463	0.123	0.463
2000	0.774	0.086	0.774
2500	0.548	0.385	0.548
3000	0.305	0.215	0.305
3500	0.346	0.253	0.346
4000	0.357	0.273	0.357
4500	0.382	0.159	0.382
5000	0.105	0.115	0.105

For practical purposes, CSR lambda was swept across an order of magnitude with <0.01 EI variation in the final outcome. No tuning is needed; $\lambda = 0.1$ is fine. The "not searched" limitation from the original Paper 44 is now a strength: CSR is genuinely insensitive to this hyperparameter in the tested range.

6.3 CSR performance

Under light intervention (rank 16, attention-only, 2000 steps), B4 achieves EI = 0.167, a 57% reduction from the 0.392 baseline, while retaining HumanEval+ 57.3% (vs. 58.5% for B0 and 57.9% for B2). The cost is 0.6 percentage points of HumanEval+ relative to B2.

Under strong intervention (rank 64, all linear, 5000 steps), B4 is the best-performing fine-tuned condition: 49.4% HumanEval+ (81/164) vs. 48.2% for B2. CSR both preserves coding performance better *and* constrains weight displacement: the SV energy of B4's weight changes is 0.001, compared to 0.006 for B2 and 1.096 for B3.

7 Reversibility

If disentanglement is a structural reorganization of the representational geometry, is it reversible? Can continued fine-tuning with a different companion recover entanglement?

We test this by taking a B3-trained Qwen-14B adapter (seed 0, post-B3 EI = 0.279) and continuing training with B2 data (code + math—the low-crosstalk companion that preserves

entanglement in fresh training). If the B3-induced disentanglement is fragile, B2 retraining should push EI back toward baseline.

7.1 Results

Table 13 reports the EI trajectory during B2 retraining.

Table 13: Reversibility test: Qwen-14B, B3→B2 sequential fine-tuning, seed 0. Disentanglement deepens rather than reverses. EI reaches 0.000 within 1000 additional steps.

Step	EI	V-purity	Domain acc
0 (post-B3)	0.279	0.654	1.000
500	0.245	0.671	0.983
1000	0.000	0.656	1.000
1500	0.000	0.563	0.883

Disentanglement is *not* reversed—it deepens. B2 training on top of B3 drives EI from 0.279 to 0.000 within 1000 steps. The model does not re-entangle. Instead, the low-crosstalk math companion reinforces the separation that B3 initiated.

The context matters. Fresh B2 training on Qwen-32B *does not* produce disentanglement—B2's mean EI fluctuates around 0.40 for the full 5000 steps (Table 4). But once B3 has partially disentangled the representation, B2 continues the process rather than counteracting it.

Table 14: Reversibility context across scales. The single valid test (14B) shows deepening, not reversal.

Scale	Pre-B3 EI	Post-B3 EI	Post-B3→B2 EI	Outcome
Qwen-14B	0.799	0.279	0.000	Deepened
Qwen-32B	0.609	0.000	(degenerate)	N/A

7.2 Interpretation

The result suggests a one-way ratchet: once the model finds a disentangled representation, subsequent training of any kind maintains or deepens the separation. The representational geometry has settled into a basin where domain-separated representations are locally optimal—continued gradient descent stays in (or deepens) that basin regardless of the data mixture.

Two caveats limit the strength of this conclusion. First, $N = 1$: only one seed at one scale. Second, the domain accuracy drops to 0.883 and V-purity drops to 0.563 by step 1500, suggesting the continued training may be degrading the probe classification structure while

maintaining zero EI. The 32B reversibility test was attempted but the result is degenerate—NaN adapter corruption produced a collapsed representation (all zeros) that is not usable—and is excluded.

8 Discussion

8.1 Scale dependence is the central finding

No existing theory predicts that disentanglement effectiveness should depend on model scale. The entanglement theorem characterizes EI for a *fixed* encoding; it does not contain a parameter-count term. Yet the data show a clean monotonic gradient: -32% at 7B, -93% at 14B, -100% at 32B.

The simplest explanation is capacity-based: larger models have more representational dimensions, and more dimensions allow the between-class covariance to reorganize into block-diagonal structure without sacrificing classification performance. At 7B, the model's activation space is too constrained—code and NL share too many directions for the training process to separate them fully. At 14B, there is enough room for near-complete separation. At 32B, the separation is total and deterministic.

This capacity interpretation is consistent with the nondegeneracy data. Qwen-32B's base model already has a dramatically higher nondegeneracy margin (1.781 vs. < 0.0002 for all other models), meaning its representation uses all available dimensions with substantial between-class variance. Paradoxically, this makes it *more* susceptible to reorganization, not less: there is enough structure in the representation for fine-tuning to concentrate it into domain-aligned subspaces.

8.2 The phase transition between 7B and 14B

The transition from partial to complete disentanglement is not a sharp threshold but a graded change. At 14B, two of three seeds reach zero EI and one stops at 0.188. At 32B, all eight seeds reach zero. The critical parameter count—where disentanglement becomes reliable and complete—lies somewhere in the 14B–32B range.

A finer scale ladder (e.g., Qwen at 20B or 24B, if such models existed) would narrow this boundary. The prediction is that there exists a critical scale N^* above which B3 fine-tuning deterministically produces complete disentanglement, and that N^* is architecture-dependent—Qwen's N^* lies in the 14B–32B range; other architectures' N^* may be higher or may not exist.

8.3 Creative writing constrains the framework

The B6 falsification is a strength, not a weakness. It identifies the boundary conditions of the crosstalk prediction framework: the crosstalk matrix measures interference, not transfer. A good companion must score well on *both* axes—low interference (preserving the target domain's representations) *and* high transfer (providing reasoning structure that helps the target domain). Math achieves both (CT = 0.088, HumanEval+ 48.2%). Creative writing achieves one and fails the other (CT = 0.0015, HumanEval+ 41.5%).

The practical heuristic is now more precise: choose companions with low measured crosstalk *among those that provide relevant inductive bias.* Crosstalk is a filter, not a ranking—it eliminates bad companions (high crosstalk = high interference) but does not distinguish among the survivors without an independent measure of domain relevance.

8.4 CSR robustness validates the approach

The <0.01 EI variation across a 10× lambda range eliminates hyperparameter sensitivity as a practical concern. CSR can be applied with a default $\lambda = 0.1$ and the 90% variance threshold without tuning. Under strong intervention, B4 (CSR) is the best-performing condition (49.4% vs. 48.2% for B2 without CSR), with 1000× less weight displacement than B3.

8.5 Irreversibility has implications for model editing

If disentanglement is a one-way ratchet, model editing techniques that assume the base model's representational geometry is preserved after fine-tuning may be invalid for fine-tuned models at scale. The representational basis has changed—directions that carried all concepts in the base model may carry only one concept after fine-tuning. Interpretability tools calibrated on base models should be re-validated on fine-tuned variants.

8.6 Entanglement-performance dissociation

Under light intervention on Qwen-32B, B2 achieves the best HumanEval+ (57.9%) with the highest EI (0.563), while B4 achieves the lowest EI (0.167) with only slightly lower HumanEval+ (57.3%). This confirms that EI measures representational *organization*, not task *quality*. A model can be highly entangled and perform well.

Under strong intervention, the dissociation partially resolves: B4 is both the least-displaced condition (SV energy 0.001) and the best-performing (49.4%). At higher intervention strength, where uncontrolled drift becomes the dominant failure mode, constraining drift provides a performance benefit as well as a structural one.

9 Limitations

Single-seed reversibility. The reversibility result is $N = 1$ (one seed, one scale). The conclusion that disentanglement is irreversible is suggestive, not established. Additional seeds at 14B and a non-degenerate 32B test are needed.

Noisy cross-architecture results at 7B. DeepSeek shows high seed variance (std = 0.210 across 2 seeds). Mistral's seed 42 shows a slight EI *increase* while seeds 0 and 1 show decreases. A third seed for each architecture would improve confidence.

Creative writing mechanism is unexplained. The B6 falsification identifies *that* low crosstalk is insufficient, but the mechanism by which creative writing actively harms coding performance—despite near-zero representational interference—is not characterized. Possibilities include: dilution of the code training signal, displacement of code-specific representations by creative structure, or incompatible optimization dynamics.

CSR lambda tested at one seed per value. The lambda sweep uses seed 0 only. The identical trajectories at $\lambda = 0.03$ and $\lambda = 0.30$ are consistent with the same seed and data order, suggesting the sweep tests lambda sensitivity but not seed×lambda interaction.

CodeLlama anti-disentanglement is unexplained. CodeLlama's EI increases from 0.874 to 1.143 (+31%) under B3. Both seeds show this effect. One possibility: CodeLlama's representations are already code-specialized, and forcing NL co-training creates destructive interference rather than separation. Testing at a larger CodeLlama scale (if available) would determine whether this is a 7B-specific effect.

Probe validity under large perturbation. We use the same factorial probes throughout training. Under strong intervention, the activation geometry may shift beyond the probes' measurement range—probe validity may degrade under large representation shifts. The V-purity and domain accuracy reported at each 500-step checkpoint serve as validity checks: stable V-purity indicates the probe infrastructure is still measuring the same quantities (Section **??**). EI measurements for the scale ladder and cross-architecture experiments use the same probe infrastructure validated on base models. Developing EI measures robust to large representational shifts is an open problem.

Single model family for the scale ladder. The scale dependence is established within the Qwen-2.5-Coder family only. Other architecture families may show different scale gradients.

Confirming the pattern in Llama, Mistral, or DeepSeek families at multiple scales would strengthen the finding.

10 Conclusion

Fine-tuning changes how a model organizes its knowledge, and the reorganization follows a scale-dependent pattern.

Within the Qwen-2.5-Coder family, B3 fine-tuning (code + NL companion) reduces entanglement intensity by 32% at 7B (0.641 ± 0.099, 3 seeds), 93% at 14B (0.063 ± 0.109, 3 seeds), and 100% at 32B (0.000 ± 0.000, 8 seeds). The transition from partial to complete disentanglement lies between 7B and 14B. At 32B, the collapse is a sharp phase transition occurring at steps 2000–3500, with temporary recoveries in 25% of seeds consistent with oscillation near the non-degeneracy boundary.

At 7B, the effect is architecture-dependent: Qwen shows 32% reduction, DeepSeek 24%, Mistral 10%, and CodeLlama shows 31% *increase*. Disentanglement at 7B depends on architecture-specific details; at 14B+, it is robust within the Qwen family.

The crosstalk matrix partially predicts companion-selection outcomes: math companions (CT = 0.088) retain 11.6 percentage points more coding performance than NL companions ($p = 0.016$). But the creative writing companion—lowest measured crosstalk at CT = 0.0015—produces the worst non-curriculum performance (41.5%, $p = 0.001$ vs. baseline), falsifying naive crosstalk minimization. Low interference is necessary but not sufficient; companions must also provide transferable inductive bias.

CSR is robust to hyperparameter choice (<0.01 EI variation across a 10× lambda range) and is the best-performing fine-tuned condition under strong intervention (49.4% HumanEval+, 81/164). Preliminary evidence ($N = 1$) suggests disentanglement may be irreversible at this scale: B3→B2 sequential fine-tuning on Qwen-14B drives EI from 0.279 to 0.000 within 1000 steps rather than recovering entanglement (needs replication).

The scale-dependent collapse has a theoretical implication: entanglement is not a geometric inevitability but a diagnostic of how the model compresses its training data. The concentration-of-measure bound holds for a given informative subspace rank r, but the model chooses r. At sufficient scale, the model can expand r to separate concepts entirely. Entanglement intensity is therefore a measurement tool—it quantifies the degree of representational compression, not a fundamental limit on representation.

What to do next: replicate the scale ladder in a non-Qwen family, test reversibility at multiple seeds and scales, and develop a two-dimensional companion-selection metric that captures both crosstalk (interference) and domain relevance (transfer) to replace the

one-dimensional crosstalk ranking that the creative writing result has falsified.

References

A. Aghajanyan, S. Gupta, and L. Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proc. ACL*, 2021.

R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient finetuning of quantized language models. In *NeurIPS*, 2023.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veres, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Kavukcuoglu, R. Pascanu, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

J. Liu, C. S. Xia, Y. Wang, and L. Zhang. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. In *NeurIPS*, 2024.

X. Ma, X. Chu, Z. Yang, Y. Lin, X. Gao, and J. Zhao. Parameter efficient quasi-orthogonal fine-tuning via Givens rotation. In *ICML*, 2024.

A. Mallya and S. Lazebnik. PackNet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, 2018.

J. McEntire. Universal entanglement in transformer activation space: Discovery, replication, and the discrimination–activation dissociation. Zenodo, 2026. doi:10.5281/zenodo.19409951.

A. Mueller, A. Lee, S. Joshi, E. S. Lubana, D. Sridhar, and P. Reizinger. From isolation to entanglement: When do interpretability methods identify and disentangle known concepts? *arXiv preprint arXiv:2512.15134*, 2025.

Z. Qiu, W. Liu, H. Feng, Y. Xue, Y. Feng, Z. Liu, D. Zhang, A. Weller, and B. Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*, 2023.

Qwen Team. Qwen2.5-Coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

W. F. Shen, X. Qiu, N. Cancedda, and N. D. Lane. Don't make it up: Preserving ignorance awareness in LLM fine-tuning. *arXiv preprint arXiv:2506.14387*, 2025.

F. Wu, J. Hu, G. Min, and S. Wang. Efficient orthogonal fine-tuning with principal subspace adaptation. *arXiv preprint arXiv:2505.11235*, 2025.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017.

R. Zhong, T. Lei, D. Yang, and J. Steinhardt. Do fine-tuning and retrieval change the activation patterns of pre-trained language models? *arXiv preprint arXiv:2510.09359*, 2025.