Entanglement-Optimal Fine-Tuning:

Crosstalk-Guided Companion Selection and Complement-Subspace Regularization for Code Models

Jeremy McEntire[*]

Abstract

Structural entanglement—the finding that every informative direction in a transformer's activation space carries all concept dimensions simultaneously—predicts that fine-tuning will produce cross-domain interference proportional to the crosstalk between training and non-training domains. We test this prediction by measuring the block-diagonal structure of the removal damage matrix across six domain pairs in Qwen 2.5-Coder-7B, then fine-tuning Qwen 2.5-Coder-32B under six conditions designed to vary the predicted interference, at two intervention strengths.

Four results. First, domain pairs differ dramatically in crosstalk: code–creative-writing is nearly block-diagonal (crosstalk 0.0015, block-diagonal ratio 0.99), while code–legal shows heavy cross-coupling (crosstalk 0.30, ratio 0.50). Second, the low-crosstalk companion (math, crosstalk 0.088) produces better code performance than the high-crosstalk companion (natural language): under strong intervention (QLoRA rank 64, all linear layers, 5,000 steps), HumanEval+ 48.2% vs. 36.6% ($p = 0.016$, two-proportion z-test). Third, complement-subspace regularization (CSR)—a geometric constraint that penalizes activation drift in non-target directions during training—is the best-performing fine-tuned condition under strong intervention (HumanEval+ 49.4%) while reducing entanglement intensity by 57% under light intervention. Fourth, curriculum ordering matters: sequential math-then-code training catastrophically degrades code performance (HumanEval+ 10.4%), while interleaved math+code training preserves it.

Cross-family replication on CodeLlama-7B and DeepSeek-Coder-6.7B shows that B3's EI collapse is Qwen-specific: both non-Qwen models maintain or increase EI under code+NL training. Probe validation using three independent probe sets (including 132 probes from real datasets) confirms the Qwen collapse is genuine geometric destruction, not a measurement artifact. These results demonstrate that the crosstalk structure of

[*] Working Paper. Correspondence: `jmc@cageandmirror.com`

the activation space predicts fine-tuning interference, that geometric regularization can mitigate it, and that the EI response to companion training is model-family-dependent.

1 Introduction

Fine-tuning large language models on domain-specific data is standard practice, but the choice of *what else* to include in the training mix is largely heuristic. Practitioners know that adding some math data helps coding models [Qwen Team, 2025] and that irrelevant data hurts, but the mechanism behind these observations is unclear. Why does math help code? Why does general natural language hurt?

The structural entanglement phenomenon [McEntire, 2026a] provides a framework for answering these questions. In Qwen 2.5-7B, every informative direction in the activation space carries all concept dimensions simultaneously. The entanglement theorem [McEntire, 2026b] formalizes this: for k concepts encoded in $d \gg k$ dimensions via multi-concept ridge regression with informative rank r and largest concept subspace dimension $d_{\max} = \max_j(m_j - 1)$, the entanglement intensity is characterized by

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

which exceeds 1 whenever $d_{\max} < r/2$—i.e., whenever no single concept dominates the informative subspace. The specialist bound (Corollary 3 of McEntire 2026b) offers the escape: since EI scales superlinearly with k (the number of simultaneously tracked concept axes), distributing concepts across specialist modules with $k_i \ll k$ reduces entanglement.

An important clarification: EI measures representational *organization*—how tangled the model's internal concept encodings are—not task *quality*. A model can be highly entangled and perform well, or minimally entangled and perform equally well. The prediction we test is not that lower EI produces better performance, but that the *crosstalk structure* of the activation space predicts which companion domains will interfere with fine-tuning. We operationalize three predictions:

1. **Crosstalk is measurable and varies by domain pair.** The removal damage matrix should show near-block-diagonal structure for "compatible" domain pairs (code + math) and heavy cross-coupling for "incompatible" pairs (code + medical).
2. **Low-crosstalk companions improve fine-tuning.** Training with a low-crosstalk companion domain should produce better target-domain performance than training with a high-crosstalk companion.
3. **Complement-subspace regularization reduces entanglement.** Penalizing ac-

tivation drift in non-target directions during training should reduce post-training EI without destroying target-domain performance.

2 Background

2.1 Structural entanglement and the specialist bound

Structural entanglement was established empirically in McEntire [2026a] and formalized in McEntire [2026b]. The key objects are:

- 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 from the classifier.
- **Entanglement intensity** $\mathrm{EI} = \sum_i \sum_{j \neq \mathrm{own}(i)} D_{ij} / \sum_i D_{i,\mathrm{own}(i)}$, the ratio of off-diagonal to diagonal damage.

The specialist bound [McEntire, 2026b, Corollary 3] states that EI scales superlinearly with k: for equal-cardinality concepts, $\mathrm{EI}_k/\mathrm{EI}_2 \geq k-1$. A specialist model tracking $k_i \ll k_{\mathrm{total}}$ concepts therefore achieves lower EI than a generalist tracking all concepts. In the symmetric case ($d_j = r/k$ for all j), the characterization $\mathrm{EI} \approx (r - d_{\max})/d_{\max}$ reduces to $\mathrm{EI} \approx k - 1$, so halving the number of tracked concepts roughly halves EI.

2.2 Block-diagonal structure in the damage matrix

If two domains A and B use non-overlapping activation subspaces, removing a direction informative for A should not damage classification of B. The 2×2 block of D corresponding to the (A, B) pair should be diagonal. We quantify this with:

Definition 1 (Symmetric crosstalk)**.** *For domains A and B with normalized damage values $D_{A \to B}$ (damage to B from removing A's directions) and $D_{B \to A}$, normalized by row-sum so that $\sum_j D_{A \to j} = 1$ (each domain's total damage budget is 1):*

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

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 (no crosstalk); 0.5 means the off-diagonal damage equals the diagonal damage.

Definition 3 (Crosstalk asymmetry)**.**

$$Asym(A, B) = 1 - \frac{\min(D_{A\to B},\ D_{B\to A})}{\max(D_{A\to B},\ D_{B\to A})}.$$

Asymmetry of 1.0 means the crosstalk is perfectly unidirectional (one cross-term is zero); 0.0 means equal bidirectional crosstalk. The metric is undefined when both cross-terms are zero (perfectly block-diagonal), reported as 1.0 by convention.

3 Experiment A: Block-Diagonal Structure Measurement

3.1 Setup

We measure the pairwise crosstalk structure of the removal damage matrix in Qwen 2.5-Coder-7B across six knowledge domains: **code**, **math**, **formal logic**, **medical**, **legal**, and **creative writing**. For each domain, we construct factorial probes varying domain (6 levels), register (4 levels: formal, informal, technical, conversational), and reasoning shape (4 levels: hierarchical, causal, constraint, evidence), yielding $6 \times 4 \times 4 = 96$ probes per domain, 480 total.

Activations are captured at the terminal layer (layer 27, hidden dimension 3,584). We fit a multi-output ridge classifier ($\alpha = 1.0$) and compute the full removal damage matrix via leave-one-out cross-validation with per-fold centering.

3.2 Results

Baseline domain classification accuracy is 95.4%, confirming the probes span the activation space adequately. Aggregate entanglement intensity is 1.175, consistent with structural entanglement at 7B scale.

Table 1 shows the pairwise crosstalk between code and each companion domain, ranked from lowest to highest.

The crosstalk structure is highly non-uniform. Code and creative writing are nearly perfectly block-diagonal (BDR = 0.991)—removing code-informative directions causes almost zero damage to creative-writing classification, and vice versa. These domains occupy nearly orthogonal activation subspaces. At the other extreme, code and legal show heavy bidirectional interference (BDR = 0.499), with the lowest asymmetry score (0.23), meaning the crosstalk flows in both directions.

Table 1: 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 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

Math and formal logic cluster together at intermediate crosstalk (0.088 and 0.092 respectively), with asymmetry 1.00—the interference is perfectly unidirectional: removing math-informative directions damages code classification, but removing code-informative directions does not damage math classification. Four of five pairs show asymmetry 1.00, meaning one of the two cross-domain damage values is exactly zero (to within measurement precision). Only code–legal shows bidirectional interference (asymmetry 0.23).

Prediction assessment. We predicted that code–math and code–formal-logic would show the lowest crosstalk. They are second and third lowest, confirming the prediction for "architecturally adjacent" domains. Creative writing shows the lowest crosstalk—nearly zero—which we did not predict. Post-hoc, this is consistent with the entanglement theorem's concept-type independence corollary: creative writing uses fundamentally different representational geometry than code, occupying a genuinely orthogonal subspace. Adjacent domains (math, formal logic) share *some* reasoning structure with code, reducing crosstalk relative to distant domains (medical, legal) but maintaining nonzero overlap. We note this interpretation is post-hoc; the prediction was partially correct in ranking but wrong about which pair would be lowest.

The full 6 × 6 crosstalk matrix reveals additional structure: medical and creative writing have zero symmetric crosstalk between each other (BDR = 1.00), and formal-logic and medical are also near-orthogonal (crosstalk 0.064). The matrix is far from symmetric—domain pair structure depends on which domain's directions are removed.

4 Experiment B: QLoRA Fine-Tuning with Varied Companions

4.1 Setup

We fine-tune Qwen 2.5-Coder-32B using QLoRA [Dettmers et al., 2023] under six conditions at two intervention strengths:

Table 2: Fine-tuning conditions. B0–B4 are tested at both light and strong intervention; B5–B6 at strong only.

Condition	Training Data	Rationale
B0: Baseline	None (base model)	Published baseline 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 (§5)
B5: Curriculum	1,000 steps math, then code	Sequential domain transfer
B6: Code+Creative	70% code, 30% creative writing	Lowest-crosstalk companion (CT = 0.0015)

We run two intervention strengths to test whether the companion-selection effects are robust:

Table 3: QLoRA configurations. Both use 4-bit NF4 with double quantization, dropout 0.05, cosine LR schedule with 3% warmup, batch size 1 with gradient accumulation 16, and gradient checkpointing.

Parameter	Light	Strong
LoRA rank (r)	16	64
LoRA α	32	128
Target modules	4 (attention)	7 (all linear)
Training steps	2,000	5,000
Learning rate	2×10^{-4}	1×10^{-4}
Max sequence length	1,024	2,048

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

Evaluation. We evaluate on HumanEval [Chen et al., 2021] and HumanEval+ [Liu et al., 2024] using the EvalPlus framework with Qwen's published evaluation protocol: raw completion mode (no chat template), prompt with trailing newline, 768 max new tokens, greedy decoding ($T = 0$), and the standard stop sequence set. All evaluations are deterministic: seven independent runs of the base model through the same pipeline return identical scores (zero variance). Each adapter is gate-verified before evaluation: model outputs with the adapter enabled must differ from base-model outputs.

Entanglement measurement. For each condition, we measure EI using the same factorial probe infrastructure as Experiment A, adapted for the 32B model with 4-bit quantization. EI is computed before and after fine-tuning. We report EI only for the light intervention, where the perturbation is small enough that the factorial probes remain valid. Under strong intervention, the activation geometry may shift beyond the probes' measurement range.

4.2 Results: Coding performance

Light intervention (pilot). Under light intervention, all conditions cluster within 1.3 percentage points of baseline on HumanEval+ (Table 4). The B2 > B3 ordering is consistent with the crosstalk prediction but the 1.2-point gap is not statistically distinguishable from zero (two-proportion $z = 0.22$, $p = 0.41$; see §4.4). The light intervention is too weak to test the companion-selection hypothesis.

Table 4: Light intervention: HumanEval results (pass@1, %). Rank 16, attention-only, 2,000 steps. B0 reproduces the published 65.9% baseline.

Condition	HumanEval	HumanEval+	Train Loss	Δ HE+ vs B0
B0: Baseline	65.9	58.5	—	—
B1: Code-only	65.2	57.3	1.15	−1.2
B2: Code+Math	65.2	57.9	0.98	−0.6
B3: Code+NL	65.2	56.7	1.25	−1.8
B4: Code+Math+CSR	64.6	57.3	0.97	−1.2

Strong intervention (main experiment). Under strong intervention (rank 64, all linear layers, 5,000 steps), the conditions separate substantially (Table 5). The same B4 > B2 > B1 > B3 ordering holds, with the B2–B3 gap widening from 1.2 to 11.6 percentage points ($z = 2.15$, $p = 0.016$). All conditions degrade from B0, which is expected: the base model is already optimized for code, and fine-tuning on smaller, lower-quality data cannot improve it. The claim is not that fine-tuning improves the model—it is that *how* one fine-tunes determines how much capability is lost.

Four observations from the strong intervention:

Table 5: Strong intervention: HumanEval results (pass@1, %). Rank 64, all linear layers, 5,000 steps. Evaluation is deterministic (greedy decoding, $T = 0$); each adapter is gate-verified before evaluation.

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

1. **B2 > B3** (HumanEval+ 48.2% vs. 36.6%, $p = 0.016$). The math companion retains 11.6 more percentage points of coding performance than the NL companion. This is a large effect: B3 loses nearly twice as much performance as B2 relative to baseline.
2. **B4 is the best fine-tuned condition** (49.4% HumanEval+), outperforming B2 by 1.2 points. CSR not only reduces entanglement (§5) but also preserves coding performance better than unrestricted fine-tuning under strong intervention.
3. **B6 falsifies naive crosstalk minimization** (41.5% HumanEval+, $p = 0.002$ vs. B0). The companion with the lowest measured crosstalk (code–creative-writing, CT = 0.0015) produces the second-worst performance, losing 17.0 percentage points. Low crosstalk does not imply good transfer (§8.4).
4. **B5 is catastrophic** (10.4% HumanEval+). Sequential math-then-code training destroys coding ability. The 1,000 math-only steps move the model into a region of weight space that 4,000 subsequent code steps cannot recover. This contrasts sharply with B2 (interleaved math+code), confirming that domain ordering, not just domain composition, determines fine-tuning outcomes.

4.3 Results: Entanglement intensity

Light intervention (pilot). Under light intervention, EI measurements remain within the probes' measurement range (Table 6).

Under light intervention, B1 and B2 *increase* EI by ∼43%, while B3 decreases it by 17% and B4 achieves the largest reduction (−57%) via CSR (§5). The ordering EI(B4) < EI(B3) < EI(B0) < EI(B1) ≈ EI(B2) shows that entanglement and task performance are partially decoupled: B2 has the highest EI *and* the best HumanEval+ among fine-tuned conditions.

Table 6: Entanglement intensity before and after fine-tuning (light intervention, single seed). Measured via removal damage on factorial probes with LOO-CV Ridge classification.

Condition	**EI (before)**	**EI (after)**	**Change**
B0: Baseline	0.392	0.392	—
B1: Code-only	0.392	0.557	+42%
B2: Code+Math	0.392	0.563	+44%
B3: Code+NL	0.392	0.325	−17%
B4: Code+Math+CSR	0.392	0.167	−57%

Strong intervention (8-seed replication). To test whether EI patterns are stable under strong intervention, we trained 8 independent seeds per condition (rank 64, all linear layers, 5,000 steps) and measured EI at each checkpoint using the same factorial probes. Table 7 reports bootstrap 95% CIs over 10,000 resamples.

Table 7: Entanglement intensity under strong intervention ($n = 8$ seeds per condition, 95% bootstrap CI on ΔEI). B4's pre-training EI differs because CSR initializes the complement-subspace projection before the first measurement.

Condition	**EI (before)**	**EI (after)**	**ΔEI**	**95% CI (Δ)**
B1: Code-only	0.622	0.345	−0.277	[−0.447, −0.088]
B2: Code+Math	0.622	0.350	−0.272	[−0.428, −0.114]
B3: Code+NL	0.622	0.000	−0.622	[−0.622, −0.622]
B4: Code+Math+CSR	0.301	0.198	−0.103	[−0.231, +0.039]

Three findings emerge from the strong-intervention replication:

- **B3 drives EI to zero with zero variance across seeds.** All 8 seeds converge to EI = 0.000 by step 3,500, indicating complete destruction of the measured entanglement structure. The EI trajectory shows a smooth collapse: 0.37 at step 500, 0.14 at step 2,000, and 0.00 from step 3,500 onward. This is consistent with NL training broadly overwriting domain-specific activation geometry.
- **B4 (CSR) has the smallest EI change**, with a 95% CI that spans zero [−0.231, +0.039]. CSR preserves the pre-training entanglement structure even under strong intervention. The ordering reverses from light intervention: B1 and B2 now *decrease* EI rather than increasing it, suggesting that under strong perturbation, code-focused training erodes rather than sharpens the discriminative structure measured by factorial probes.
- **B1 and B2 are indistinguishable** (ΔEI of −0.277 vs. −0.272, overlapping CIs). The math companion does not significantly alter how much entanglement structure is lost

relative to code-only training. Its performance benefit (Table 5) arises from a different mechanism than EI preservation.

The reversal between light and strong intervention—B1/B2 increase EI under light perturbation but decrease it under strong—suggests that moderate fine-tuning sharpens domain directions (increasing off-diagonal damage ratios) while aggressive fine-tuning overwhelms them. This is consistent with the interpretation that EI measures representational organization, not task quality: the same conditions that preserve coding performance (B4 best, B3 worst) also preserve entanglement structure, but only under strong enough intervention that uncontrolled drift is the dominant failure mode.

4.4 The specialist bound at scale

The block-diagonal experiment measured EI = 1.175 in 7B with 6 domain axes. The QLoRA experiment measures EI = 0.392 in 32B with 6 domain axes, a 3× reduction. This is *consistent* with the specialist bound but not a clean test: the two measurements differ in model size (7B vs 32B), training data composition, architecture depth, and quantization (full precision vs 4-bit). Any of these factors could contribute to the EI difference. A controlled test would require same-family models at multiple scales with identical measurement protocols. We report this comparison as suggestive, not confirmatory.

Statistical analysis. All HumanEval evaluations use greedy decoding ($T = 0$), producing deterministic outcomes: each problem either passes or fails with probability 1. We treat the 164 HumanEval+ problems as a finite sample from the population of possible coding tasks. The two-proportion z-test compares conditions:

For the strong intervention B2 vs. B3 gap (48.2% vs. 36.6%):

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

where $\hat{p} = (79 + 60)/(2 \times 164) = 0.424$ is the pooled proportion. The 95% confidence interval for the difference is $[1.1, 22.1]$ percentage points.

For the light intervention B2 vs. B3 gap (57.9% vs. 56.7%): $z = 0.22$, $p = 0.41$. Not statistically significant.

The strong intervention B4 vs. B3 gap (49.4% vs. 36.6%): $z = 2.34$, $p = 0.010$. The strong intervention B1 vs. B3 gap (47.6% vs. 36.6%): $z = 2.02$, $p = 0.022$. Under strong intervention, NL is significantly worse than all other companion choices.

The B6 vs. B0 gap (41.5% vs. 58.5%): $z = -3.08$, $p = 0.001$, 95% CI $[-27.7, -6.3]$ pp. Creative writing as companion *significantly harms* coding performance despite near-zero crosstalk (CT = 0.0015).

Statistical caveat. The two-proportion z-test treats the 164 HumanEval+ problems as independent Bernoulli draws from a population of coding tasks. This is standard practice in the evaluation literature but the "population" is notional—HumanEval is a fixed benchmark, not a random sample. The test quantifies how surprising the observed difference would be under a null model of equal capability, not the probability that the difference generalizes to arbitrary coding tasks. We report it as the conventional measure while noting this limitation.

5 Experiment C: Complement-Subspace Regularization

5.1 Mechanism

Standard fine-tuning updates weights in all directions, including those that serve non-target domains. We introduce a regularization term that 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 $d = 5{,}120$ for 32B).
2. Compute the target-domain subspace $S_{\text{target}} \in \mathbb{R}^{d \times p}$ via truncated SVD of the $n_{\text{target}} \times d$ matrix of target-domain activation vectors, retaining the top p singular vectors that capture 90% of the variance. In our experiments, $p = 47$ for the code domain (out of $d = 5{,}120$), so the complement subspace has dimension $d - p = 5{,}073$.
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 the regularization loss:

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

where $\lambda = 0.1$ is the regularization strength (fixed, not searched). The 90% variance threshold for p was chosen to include the primary representational subspace while leaving the complement large enough to constrain drift meaningfully. Sensitivity to this threshold is not explored in this work.

Relationship to EWC. Elastic Weight Consolidation [Kirkpatrick et al., 2017] penalizes weight changes weighted by Fisher information. CSR operates in *activation space* rather than weight space, targeting the geometric structure that the entanglement theorem describes. EWC asks "which weights matter?"; CSR asks "which activation directions are mine?"

Relationship to inference-time stochastic resonance. Stochastic resonance techniques inject noise through domain projections at inference time to amplify weak signals. CSR is fundamentally different: a geometric constraint during *training*, not noise injection during inference. It shares the projection-matrix machinery but applies it to a different problem (preventing drift vs. amplifying signal).

5.2 Results

Light intervention. B4 achieves EI = 0.167, a 57% reduction from the 0.392 baseline, while retaining 57.3% HumanEval+ (vs. 58.5% for B0 and 57.9% for B2). The cost is 0.6 points of HumanEval+ relative to B2 (no regularization, same data).

Strong intervention. B4 is the best-performing fine-tuned condition: 49.4% HumanEval+ vs. 48.2% for B2 (same data, no CSR). Under light intervention, CSR slightly underperformed B2 in coding accuracy (57.3% vs. 57.9%) while dramatically reducing EI. Under strong intervention, CSR both preserves coding performance better *and* constrains weight displacement (Table 8).

Table 8: Singular value energy of weight changes under strong intervention. For B1–B4, values are means over 8 seeds with 95% bootstrap CIs. B5 and B6 are single-seed. Lower values indicate more focused weight updates.

Condition	SV Energy	95% CI	HE+
B4: Code+Math+CSR	0.032	[0.000, 0.092]	49.4
B1: Code-only	0.001	[0.000, 0.002]	47.6
B2: Code+Math	0.012	[0.000, 0.034]	48.2
B6: Code+Creative	0.041	—	41.5
B5: Curriculum	0.646	—	10.4
B3: Code+NL	1.514	[0.317, 3.160]	36.6

The 8-seed replication confirms the qualitative pattern from the pilot: B3 (NL companion) produces > 1,000× more weight displacement than code-focused conditions, with the CI excluding the B1/B2/B4 ranges entirely. B4's SV energy CI overlaps with B2, consistent with the CSR penalty constraining but not eliminating complement-subspace drift. SV energy remains strongly anticorrelated with coding performance across conditions. B6 reinforces this pattern: its moderate SV energy (0.041) produces poor coding outcomes, confirming that focused weight changes are necessary but not sufficient—the companion must also provide transferable inductive bias.

CSR loss trajectory. The $\mathcal{L}_{\text{CSR}}$ loss trajectory (Figure 1) shows three phases:

1. **Rapid growth** (steps 0–500): $\mathcal{L}_{\text{CSR}}$ rises from 0.01 to 6.7 as the model begins learning and activations diverge from the base model.
2. **Deceleration** (steps 500–1350): Growth rate decreases, reaching 14.6 at step 1350, as the model finds directions that serve code without drifting excessively in the complement.
3. **Plateau** (steps 1350–2000): $\mathcal{L}_{\text{CSR}}$ stabilizes near 15.0, indicating convergence to a region balancing task loss and complement-subspace constraint.

CSR loss trajectory over 2,000 training steps (light intervention).

Step 50: 0.01 → Step 500: 6.7 → Step 1000: 11.0 → Step 1350: 14.6 → Step 2000: 15.0

Three-phase structure: rapid growth, deceleration, plateau.

Figure 1: CSR regularization loss $\mathcal{L}_{\text{CSR}}$ during B4 training. The plateau indicates convergence to a task-optimal, drift-minimal region. (Data summary; rendered plot for submission version.)

6 Experiment D: Cross-Family Replication

The Qwen results raise a natural question: does the B3 EI collapse generalize across model families, or is it specific to the Qwen architecture? We replicate the B2 vs. B3 comparison on two additional code-specialized model families: CodeLlama-7B [Rozière et al., 2023] and DeepSeek-Coder-6.7B [Guo et al., 2024].

6.1 Setup

We train B2 (code+math) and B3 (code+NL) on each model using identical QLoRA configuration to the strong-intervention Qwen experiment: rank 64, $\alpha = 128$, all 7 linear modules, 5,000 steps, learning rate 1×10^{-4}, cosine schedule. Effective batch size is 16 (train batch 2 × gradient accumulation 8). CodeLlama runs at two seeds per condition; DeepSeek at one seed per condition. EI is measured using the same 60 factorial probes at 500-step intervals.

6.2 Results

Three findings:

1. **B3 does not collapse EI on either non-Qwen model.** CodeLlama B3 *increases* EI (from 0.874 to 1.093–1.192), and DeepSeek B3 *decreases* EI only modestly (from 1.376 to 1.196). Neither approaches the EI = 0 collapse observed on Qwen. The B3 EI collapse is model-family-specific.

Table 9: Cross-family EI replication. CodeLlama reports mean over 2 seeds; DeepSeek is single-seed.

Model	**Condition**	**EI (before)**	**EI (after)**	**ΔEI**
CodeLlama-7B (s0)	B2: Code+Math	0.874	1.033	+0.159
CodeLlama-7B (s0)	B3: Code+NL	0.874	1.093	+0.220
CodeLlama-7B (s1)	B2: Code+Math	0.874	1.077	+0.203
CodeLlama-7B (s1)	B3: Code+NL	0.874	1.192	+0.319
DeepSeek-6.7B (s0)	B2: Code+Math	1.376	1.078	−0.298
DeepSeek-6.7B (s0)	B3: Code+NL	1.376	1.196	−0.180
Qwen-32B (8 seeds)	B2: Code+Math	0.622	0.350	−0.272
Qwen-32B (8 seeds)	B3: Code+NL	0.622	0.000	−0.622

2. **CodeLlama: B3 increases EI more than B2.** Across both seeds, NL training produces larger EI gains than math training (+0.220/+0.319 vs. +0.159/+0.203). This is the opposite of the Qwen pattern, where B3 destroys EI. At 7B scale with Llama-2 architecture, NL training deepens entanglement rather than destroying it.
3. **Baseline EI varies substantially across families.** DeepSeek starts at EI = 1.376, CodeLlama at 0.874, and Qwen-32B at 0.622. DeepSeek's representations are already deeply entangled before fine-tuning; both conditions reduce EI toward ∼1.0. These differences may reflect architecture, scale, or pre-training data composition.

The cross-family results transform the interpretation of the Qwen B3 finding: rather than a general property of NL companion training, the EI collapse appears to be an interaction between the Qwen architecture (or scale, or pre-training distribution) and NL fine-tuning data. This motivates the probe validation experiment below.

7 Experiment E: Probe Validation of B3 Collapse

A potential artifact could explain the Qwen B3 EI collapse: the 60 hand-crafted factorial probes were designed for the base model's representational space. If B3 training shifts the activation geometry sufficiently, the probes might become misaligned with the new representational structure, producing EI = 0 as a measurement artifact rather than genuine geometric destruction. We test this by evaluating B3 with three independent probe sets.

7.1 Setup

We load the Qwen-2.5-Coder-32B base model and the B3 adapter (seed 5, from the 8-seed replication) in 4-bit quantization, and measure EI on three probe sets:

- **ORIGINAL**: The same 60 hand-crafted probes used throughout this paper (3 domains × 4 registers × 5 shapes × 1 text each, minus incomplete cells).
- **EXPANDED**: 132 probes drawn from real datasets—code probes from `code_search_net`, math probes from `gsm8k`, and medical probes from `pubmed_qa`—assigned to the factorial structure by content type.
- **DIVERSE**: 60 new hand-crafted probes with entirely different text content from the original set, maintaining the same factorial structure (3 domains × 2 registers × 2 shapes).

If the B3 EI collapse is a probe-alignment artifact, it should disappear on the expanded and/or diverse probe sets while remaining on the original set.

7.2 Results

Table 10: Probe validation: EI across three independent probe sets on base and B3 models.

Probe Set	Model	EI	Domain	Register	Shape	SV Energy
Original (60)	BASE	0.085	1.000	0.850	0.933	0.001
Original (60)	B3	0.000	0.933	0.817	0.783	13.231
Expanded (132)	BASE	0.773	1.000	0.864	0.796	0.056
Expanded (132)	B3	0.000	0.962	0.712	0.705	0.115
Diverse (60)	BASE	0.718	1.000	0.850	0.850	0.002
Diverse (60)	B3	0.000	0.900	0.750	0.833	0.196

B3 drives EI to 0.000 on all three probe sets. The collapse is not a measurement artifact. Even 132 probes drawn from real datasets (code_search_net, gsm8k, pubmed_qa) show complete entanglement destruction under B3. The BASE model produces substantial EI on the expanded (0.773) and diverse (0.718) probe sets, confirming they measure real geometric structure.

Three additional observations:

1. **Domain classification survives** even as entanglement is destroyed: B3 achieves 0.90–0.96 domain accuracy across all probe sets. The model can still distinguish domains; it is the *cross-domain coupling*—the entanglement—that is selectively eliminated.
2. **Register and shape accuracy degrade** under B3 (from 0.85/0.85–0.93 to 0.71–0.83), indicating that NL training damages finer-grained structural representations beyond entanglement.
3. **The original 60 probes underestimate base EI** (0.085 vs. 0.718–0.773 on the larger sets). This suggests the original probe set, while adequate for detecting the B3

collapse, was suboptimal for measuring absolute EI magnitude.

Combined with the cross-family results (§6), the probe validation establishes that: (a) the Qwen B3 EI collapse is genuine geometric destruction, not a measurement artifact, and (b) this destruction is model-family-specific, not a universal property of NL companion training.

8 Discussion

8.1 What the crosstalk matrix tells us

The block-diagonal experiment reveals that domain pairs differ in crosstalk by two orders of magnitude (0.0015 for code–creative-writing vs. 0.30 for code–legal). The crosstalk matrix provides fine-grained structure beyond the aggregate EI: *which* domains interfere with *which*, and by how much.

Math is a good companion for code not because it is orthogonal (crosstalk is 0.088) but because it provides reasoning structure that transfers to code while maintaining moderate separation. The practical heuristic is: choose companions that provide useful inductive bias at low measured crosstalk.

8.2 Confounds and alternative explanations

The companion-selection experiment has inherent confounds. B1 (100% code), B2 (70% code + 30% math), and B3 (70% code + 30% NL) differ in total code data seen, companion data distribution, and vocabulary. The B2 > B3 performance difference could reflect:

- Math providing transferable reasoning skills (simple transfer learning)
- NL diluting the code signal more than math does
- Dataset-specific effects (GSM8K vs. Wikipedia quality differences)

The crosstalk explanation is consistent with the data but is not the only explanation. What the data *do* establish is that companion choice matters substantially under strong intervention (11.6pp gap), that the ordering is robust across two intervention strengths, and that SV energy provides a mechanistic correlate (B3 spreads weight changes 1,000× more broadly than B4). Disentangling the crosstalk mechanism from general transfer-learning effects would require experiments with companion domains matched on data statistics but varying in measured crosstalk—an important direction for future work.

8.3 Entanglement-performance dissociation

Under light intervention, 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%).

EI measures representational *organization*, not representational *quality*. A model can be highly entangled and perform well, or minimally entangled and perform nearly as well.

Under strong intervention, the dissociation partially resolves. The 8-seed replication shows that B4 (CSR) preserves EI better than any other condition (ΔEI = -0.103, CI spanning zero), while B3 (NL) destroys EI entirely (ΔEI = -0.622, zero variance). This ordering—B4 best, B3 worst—exactly mirrors the coding performance ordering. At higher intervention strength, where uncontrolled drift dominates, EI preservation and task performance become aligned.

8.4 Limitations

Creative-writing companion falsifies naive crosstalk minimization. Code–creative-writing has the lowest crosstalk of any pair (CT = 0.0015, BDR = 0.991). A strict reading of the crosstalk-minimization hypothesis predicts this should be the best companion. The B6 (code + creative-writing) condition, trained under strong intervention and gate-verified, achieves HumanEval pass@1 = 45.1% and HumanEval+ pass@1 = 41.5%—a catastrophic 20-point drop from baseline (65.9% / 58.5%) and the worst result of any condition. SV energy for B6 (0.041) falls between B1 (0.014) and B5 (0.646), indicating moderate weight displacement.

This result falsifies naive crosstalk minimization: the companion with the *lowest* measured crosstalk produces the *worst* coding performance. The correct reading is that low crosstalk is necessary but not sufficient. A good companion must provide useful *inductive bias* for the target domain. Math shares formal reasoning structure with code (crosstalk 0.088, HumanEval+ 57.9%); creative writing shares no such structure (crosstalk 0.0015, HumanEval+ 41.5%). The crosstalk matrix measures *interference*, not *transfer*. Low interference permits low collateral damage, but the companion's training signal must also be *helpful*—and creative writing actively displaces coding representations without providing compensating structure.

Single model family for Experiments A–C. The crosstalk measurement and CSR experiments (Experiments A–C) use only Qwen 2.5 models. The cross-family replication (Experiment D, §6) extends the B2 vs. B3 comparison to CodeLlama-7B and DeepSeek-Coder-6.7B, but does not replicate the crosstalk matrix or CSR experiments. Full cross-architecture validation of the crosstalk structure is needed.

CSR hyperparameters were not searched. We used $\lambda = 0.1$ and a 90% variance threshold for the target subspace without searching over alternatives. The optimal values likely depend on the training budget, LoRA rank, and target domain.

EI measurement under strong intervention. Under strong intervention, fine-tuning shifts the activation geometry far enough that factorial probes—designed for the base model's representational structure—may no longer measure the same quantities. The 8-seed replication (Table 7) shows that EI patterns are highly stable across seeds under strong intervention (B3 converges to EI = 0 with zero variance), suggesting the measurements capture a real geometric phenomenon. However, the absolute EI values under strong vs. light intervention are not directly comparable, and probe validity at large representational shifts remains an open question.

EvalPlus in containers. HumanEval+ evaluation required monkey-patching `resource.setrlimit` for container compatibility. We verified that B0 reproduces the published 65.9% baseline, but subtle evaluation differences cannot be ruled out.

B3 EI collapse is Qwen-specific. Under strong intervention on Qwen-32B, B3 drives EI to zero across all 8 seeds—a complete collapse of the measured entanglement structure. Probe validation (§7) confirms this is genuine geometric destruction: EI = 0 across three independent probe sets (60 original, 132 real-data, 60 diverse). However, cross-family replication (§6) shows this collapse does not occur on CodeLlama-7B (where B3 *increases* EI) or DeepSeek-Coder-6.7B (where B3 only modestly decreases EI). The collapse appears to be an interaction between the Qwen architecture (or scale, or pre-training distribution) and NL fine-tuning—not a universal property of NL companion training. Identifying which factor drives this interaction is a direction for future work.

9 Related Work

Multi-task learning and auxiliary tasks. The benefit of training with related auxiliary tasks is well established [Caruana, 1997, Ruder, 2017]. Our contribution is to provide a *measurement* (the crosstalk matrix) that predicts which auxiliary tasks will help and which will hurt, grounded in the geometry of the model's activation space.

Regularization for continual learning. EWC [Kirkpatrick et al., 2017], SI [Zenke et al., 2017], and PackNet [Mallya and Lazebnik, 2018] address catastrophic forgetting by constraining weight updates. Orthogonal Fine-Tuning [OFT; Qiu et al., 2023] constrains weight updates to orthogonal transformations, preserving pairwise angles between neuron vectors; qGOFT [Ma et al., 2024] reduces OFT's parameter cost from $O(d^2)$ to $O(d)$ via Givens rotations, and PSOFT [Wu et al., 2025] restricts orthogonal transformations to the principal subspace of pre-trained weights. SEAT [Shen et al., 2025] constrains activation drift during fine-tuning via sparse tuning and KL-divergence regularization, targeting epistemic uncertainty preservation. Our CSR differs from these methods in that it uses the entanglement structure measured by factorial probes to determine which activation-space directions to protect—information that weight-space methods (OFT, qGOFT, PSOFT) do not incorporate and that activation-drift methods (SEAT) do not target.

LoRA and parameter-efficient fine-tuning. LoRA [Hu et al., 2021] and QLoRA [Dettmers et al., 2023] reduce fine-tuning cost by learning low-rank updates. Our work is orthogonal: we do not propose a new PEFT method but rather a principled way to choose what data to train on and how to regularize the training, applicable to any PEFT approach.

Linear probing and concept geometry. The activation geometry program [McEntire, 2026a,b] established that linear probes succeed at classification but fail to identify concept-pure directions. The present work takes the first step from measurement to intervention: using the entanglement structure to guide training decisions.

10 Conclusion

Seven findings. First, the crosstalk matrix derived from the removal damage matrix provides a quantitative, per-domain-pair measure of activation-space interference. Domain pairs range from nearly orthogonal (code–creative-writing, BDR = 0.99) to heavily coupled (code–legal, BDR = 0.50). Second, under strong QLoRA intervention, a math companion retains 11.6 percentage points more coding performance than an NL companion ($p = 0.016$), with the same ordering reproduced at both intervention strengths. Third, complement-subspace regularization reduces entanglement intensity by 57% (light intervention) and is the best-performing fine-tuned condition under strong intervention (49.4% HumanEval+). Fourth, curriculum ordering matters: sequential math-then-code training catastrophically degrades code performance (10.4% HumanEval+), while interleaved training preserves it. Fifth, the 8-seed replication under strong intervention reveals that NL training drives EI to zero with zero cross-seed variance on Qwen-32B—a complete collapse of measured entanglement structure—while CSR preserves it (ΔEI CI spanning zero).

Sixth, cross-family replication on CodeLlama-7B and DeepSeek-Coder-6.7B shows that the B3 EI collapse is Qwen-specific: CodeLlama B3 *increases* EI (from 0.874 to 1.093–1.192), and DeepSeek B3 decreases it only modestly (from 1.376 to 1.196). Seventh, probe validation using three independent probe sets (including 132 probes from real datasets) confirms that the Qwen B3 collapse is genuine geometric destruction, not a measurement artifact: EI = 0.000 across all probe sets, while domain classification accuracy remains 0.90–0.96.

The companion-selection effect is consistent with the crosstalk prediction but is confounded with general transfer-learning effects; disentangling these requires further experiments. Under light intervention, high EI can coexist with good performance (B2), suggesting EI captures representational organization, not task quality. Under strong intervention this dissociation resolves: EI preservation and coding performance become aligned, and the conditions separate with non-overlapping bootstrap confidence intervals.

These results demonstrate that the activation-space geometry measured by the crosstalk matrix and EI correlates with fine-tuning outcomes, and that geometric regularization (CSR) can mitigate interference. The Qwen-specific B3 collapse—confirmed genuine by probe validation but absent in other model families—points to an interaction between architecture and NL training that merits further investigation. Whether crosstalk is the causal mechanism, or a correlate of simpler domain-similarity effects, remains open.

References

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.

D. Guo, Q. Zhu, D. Yang, Z. Xie, K. Dong, W. Zhang, G. Chen, X. Bi, Y. Wu, Y. K. Li, F. Luo, Y. Xiong, and W. Liang. DeepSeek-Coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez, J. Rapin, A. Kozhevnikov, I. Evtimov, J. Bitton, M. Bhatt, C. C. Ferber, A. Grattafiori, W. Xiong, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve. Code Llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

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.

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

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.

J. McEntire. Eight experiments on why every direction carries every concept. Zenodo, 2026. doi:10.5281/zenodo.18880969.

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

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.