Constellation-Indexed Model Composition: Query-Driven Parameter Mixing via Activation Fingerprints

Jeremy McEntire

Abstract

We introduce **Constellation-Indexed Model Composition**, a method for dynamically composing specialist language models at the parameter level using activation fingerprints and orthogonal decomposition. Given a query, we (1) capture its activation fingerprint in the generalist model, (2) project it onto Gram-Schmidt-orthogonalized domain centroids to obtain per-domain relevance weights, and (3) compose specialist parameters through node-level weighted mixing where each hidden dimension receives its own specialist mixture.

Two methodological contributions prove essential. First, **generalist-space indexing**: building the constellation index from generalist (not specialist) activations eliminates generalist-specialist alignment failures and transforms single-domain composition from broken to near-perfect (Qwen 0.5B: 20.6% → 98.1% win rate vs. generalist). Second, **Gram-Schmidt orthogonalization**: raw domain centroids exhibit mean cosine similarity 0.91–0.97 (medical-legal: 0.988–0.993), making naïve projection meaningless; orthogonalization reduces this to ∼0, enabling clean decomposition.

We evaluate across two architecture families and five scales—GPT-2 (124M), Qwen 2.5-0.5B (494M), Qwen 2.5-1.5B (1.5B), Qwen 2.5-3B (3B), and Qwen 2.5-7B (7B)—with four domain specialists (medical, legal, code, science). All five models achieve ≥93% cross-domain win rate vs. the generalist. Against task arithmetic (**?**), we observe a **peaked non-monotonic scale dependence**: constellation composition dominates at 124M (98.7% win rate vs. TA), task arithmetic briefly dominates at 0.5B (23.3%), constellation recovers through 1.5B (64.0%) and peaks at 3B (92.0%), then *declines* at 7B (60.7%). Within the Qwen architecture family, domain centroid collinearity increases monotonically with scale (0.906 → 0.934 → 0.969 → 0.973), and constellation's advantage over task arithmetic tracks collinearity from 0.5B through 3B (Spearman $\rho = 1.0$, $n = 3$), but *diverges* at 7B: collinearity continues rising while constellation's advantage falls from 92.0% to 60.7%. This reveals a **collinearity sweet spot**: constellation composition dominates when collinearity is high enough to make task arithmetic destructive (∼0.93–0.97), but at very high collinearity (>0.97) the orthogonal residuals fall below the effective precision threshold. However, a **stochastic resonance** experiment shows this signal attenuation is recoverable: at 7B, injecting Gaussian noise at $\sigma^* = 0.020\|\bar{\mathbf{c}}\|$ before orthogonalization lifts composition from 8.7% to 99.3% vs. task arithmetic (+90.7 points), while noise produces null effects at all four other scales—confirming that noise-assisted decomposition is a targeted rescue for catastrophic collinearity collapse.

The rank ordering of domain similarities is **architecture-invariant**: code is always the most distinct domain, medical-legal always the most entangled, across all five models (Spearman $\rho = 1.0$ across all four Qwen scales; $\rho = 0.75$ between GPT-2 and Qwen). This establishes that activation-space geometry is data-determined, not architecture-determined: a practitioner can diagnose which domains will be easy or hard to separate on a small, cheap model, and the answer will generalize to larger models without retraining.

1 Introduction

Domain specialization in language models follows a well-established pattern: fine-tune separate expert models for distinct tasks, then route queries to the appropriate specialist at inference time (**??**). This approach has produced impressive results but faces a fundamental limitation: queries rarely fall cleanly into a single domain. A question about medical malpractice law requires both medical and legal expertise; a query about bioinformatics pipelines requires both biology and programming knowledge. No single specialist suffices, and routing to one discards the expertise of others.

Existing approaches to combining specialist knowledge each impose significant constraints. Mixture of Experts (MoE) architectures (**??**) learn routing functions during training but require architectural modification and fixed expert allocation. Model merging methods (**??**) average specialist weights statically, producing a single merged model regardless of the query. Adapter composition approaches (**??**) operate in low-rank subspaces, limiting expressivity.

We propose a different approach: *query-driven, full-rank parameter composition* guided by activation fingerprints. The generalist model's own activations in response to a query serve as an index into a library of specialists, determining not just *which* specialists are relevant but *how* to combine their parameters at each hidden dimension. This requires no architectural modification, operates at full rank, and adapts the composed model to each query individually.

The central contributions of this paper are methodological and analytical. Methodologically, we identify two prerequisites for effective decomposition-based composition: **generalist-space indexing** (building the constellation index from generalist activations, eliminating alignment failures that plagued specialist-space indexing) and **Gram-Schmidt orthogonalization** of domain centroids (without which the raw centroids are too collinear for meaningful projection—mean cosine similarity 0.91–0.97 before orthogonalization, $\sim$0 after).

Analytically, the five-scale evaluation reveals structure that would be invisible at a single scale. The **peaked non-monotonic relationship** between model scale and task arithmetic effectiveness—constellation composition's advantage rises, peaks at 3B, then declines at 7B—provides a geometric explanation for when static merging succeeds and when query-adaptive composition is needed. The **architecture invariance** of domain similarity rankings—the same rank order across GPT-2 and all four Qwen variants—demonstrates that activation-space geometry is determined by training data, not model architecture.

This work builds on the activation fingerprinting framework introduced in our companion paper (**?**), which uses activation-space cosine similarity as a regime detection mechanism for speculative weight prediction during training. Here, we repurpose the same fingerprinting infrastructure for indexing specialist models and guiding parameter composition.

Contributions.

1. **Generalist-space indexing with Gram-Schmidt orthogonalization**: Building the constellation index from generalist (not specialist) activations and orthogonalizing domain centroids before decomposition. These two fixes transform single-domain composition from broken to near-perfect (Qwen 0.5B: 20.6% $\rightarrow$ 98.1%) and are validated across two architectures and five scales (124M–7B).

2. **Collinearity sweet spot**: Constellation composition dominates task arithmetic at 124M (98.7%), loses at 0.5B (23.3%), recovers through 1.5B (64.0%) to a peak at 3B (92.0%), then *declines* at 7B (60.7%). Domain centroid collinearity increases monotonically with scale (0.906 $\rightarrow$ 0.973), but constellation's advantage peaks and then falls—the inverted-U is the finding, not the three-point correlation. It reveals an empirically bounded operating range ($\sim$0.93–0.97 collinearity) where orthogonal decomposition maximally outperforms uniform vector addition.

3. **Architecture-invariant domain geometry**: The rank ordering of domain similarities (code most distinct, medical-legal most entangled) is perfectly preserved across all four Qwen scales ($\rho = 1.0$) and approximately preserved across architecture families ($\rho = 0.75$ GPT-2 vs. Qwen), establishing that activation-space structure is data-determined.

4. **Decomposition sparsity analysis**: Within the Qwen family, composition weight sparsity increases monotonically with scale (Gini 0.40 $\rightarrow$ 0.47 $\rightarrow$ 0.55 $\rightarrow$ 0.55), tracking re-emergent domain separability as the representation space grows.

2 Related Work

Mixture of Experts. Sparse MoE architectures (**???**) partition model capacity across expert subnetworks with learned routing. Our approach differs in that composition occurs entirely in weight space (no routing network), operates at full rank (not sparse experts), and adapts per-query without training a router.

Model Merging. Task arithmetic (**?**) and model soups (**?**) combine fine-tuned models through static weight operations. DARE (**?**) improves merging via random drop-and-rescale of task vectors. TIES (**?**) resolves interference via sign consensus. These methods produce a single merged model applied uniformly to all queries. Our approach composes *dynamically* per query, with different weight mixtures for different inputs. We compare directly against task arithmetic as a strong baseline (Section **??**).

Adapter Composition. LoRAHub (**?**) dynamically composes LoRA adapters via gradient-free optimization of mixing coefficients. Ostapenko et al. (**?**) build reusable LoRA libraries with routing mechanisms. While computationally efficient, the low-rank constraint limits the expressivity of the composed model. Our approach operates at full rank, mixing complete specialist parameters.

Activation Analysis. Representational similarity analysis (**?**) and CKA (**?**) measure similarity between neural network representations. Our activation fingerprinting is related but serves a functional purpose: indexing specialist models and generating per-node composition weights, not just measuring similarity.

3 Method

3.1 Overview

The system has three phases: (1) offline construction of a specialist constellation with an activation-based index, (2) online retrieval of relevant specialists given a query, and (3) node-level parameter composition producing a query-specific model.

3.2 Constellation Construction

Given a generalist model $\mathcal{G}$ (a pre-trained or partially trained checkpoint), we fine-tune N domain specialists $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$ by continuing training on domain-specific data from the same checkpoint. Each specialist shares the same architecture as the generalist, differing only in parameter values.

To build the constellation index, we define a probe set $\mathcal{P} = \{p_1, \ldots, p_M\}$ of M domain-specific inputs, with M/D probes per domain (where D is the number of domains).

Generalist-space indexing. A critical design choice is *which model* generates the index fingerprints. An earlier version of this system built the index from specialist activations: each probe was forwarded through each specialist, and the resulting fingerprints were stored. This creates an alignment problem: retrieval queries are generalist fingerprints, but the index contains specialist fingerprints. When the generalist represents a domain differently than the specialist (e.g., medical content: generalist-specialist centroid similarity as low as 0.736), correct retrieval becomes impossible regardless of index quality.

We resolve this by building the index entirely from *generalist* fingerprints. For each probe p_j, we forward it through the generalist $\mathcal{G}$ and extract the mean-pooled final hidden state:

$$\mathbf{h}_j = \text{MeanPool}\left(\mathcal{G}(p_j)_{\text{last}}\right) \in \mathbb{R}^d \tag{1}$$

The domain centroid for domain k is:

$$\mathbf{c}_k = \frac{1}{|\mathcal{P}_k|} \sum_{p_j \in \mathcal{P}_k} \mathbf{h}_j \tag{2}$$

Since both queries and index entries are now generalist fingerprints, the alignment condition is satisfied by construction.

Gram-Schmidt orthogonalization. The raw domain centroids $\{\mathbf{c}_k\}$ are highly collinear. Across all five model scales, the mean pairwise cosine similarity of domain centroids ranges from 0.91 to 0.97, with medical-legal reaching 0.988–0.993. Projecting a query onto nearly-parallel axes produces coefficients dominated by noise rather than genuine domain affinity.

We apply Gram-Schmidt orthogonalization to the domain centroids, producing an orthonormal basis $\{\hat{\mathbf{e}}_k\}$:

$$\hat{\mathbf{e}}_k = \text{normalize}\left(\mathbf{c}_k - \sum_{j<k} (\mathbf{c}_k \cdot \hat{\mathbf{e}}_j)\, \hat{\mathbf{e}}_j\right) \tag{3}$$

Post-orthogonalization, mean pairwise similarity drops to $\sim 10^{-9}$ (numerically zero). The query's projection onto this orthogonal basis yields clean per-domain coefficients that directly serve as composition weights.

3.3 Orthogonal Decomposition

Given a query q, we forward it through the generalist model and extract its fingerprint:

$$\mathbf{g}_q = \text{MeanPool}\left(\mathcal{G}(q)_{\text{last}}\right) \in \mathbb{R}^d \tag{4}$$

The per-domain relevance weight is the projection of the query fingerprint onto the orthogonalized domain basis:

$$\alpha_k = \mathbf{g}_q \cdot \hat{\mathbf{e}}_k \tag{5}$$

These coefficients are positive when the query genuinely aligns with domain k, and directly encode domain relevance without requiring a separate retrieval step. The top-K domains by $|\alpha_k|$ determine which specialists contribute to composition.

3.4 Node-Level Composition

This is the core contribution. For each selected specialist $\mathcal{S}_i$, we compute a per-node *activation signature*:

$$\mathbf{a}_i = \text{MeanPool}\left(|\mathcal{S}_i(q)_{\text{last}}|\right) \in \mathbb{R}^d \tag{6}$$

where the absolute value and mean are taken across the sequence dimension, producing per-hidden-dimension activation magnitudes.

The raw composition weight for specialist i at hidden dimension j is:

$$w_{i,j}^{\text{raw}} = \max(c_i, 0) \cdot a_{i,j} \tag{7}$$

where c_i is the correlation score from retrieval and $a_{i,j}$ is the activation magnitude at dimension j. Clamping c_i at zero ensures that negatively correlated specialists receive zero weight.

Two-stage normalization. We normalize weights per-node across specialists:

$$w_{i,j} = \frac{w_{i,j}^{\text{raw}}}{\sum_{i'} w_{i',j}^{\text{raw}} + \epsilon} \tag{8}$$

This produces a weight matrix $\mathbf{W} \in \mathbb{R}^{K \times d}$ where each column (hidden dimension) sums to 1, and each specialist's contribution varies across dimensions. If all weights for a particular dimension are zero (all specialists have zero activation there), we fall back to uniform weights $w_{i,j} = 1/K$.

The composed parameters for the final L transformer layers are:

$$\theta^{\text{composed}} = \sum_{i=1}^{K} \mathbf{W}_i \odot \theta_{\mathcal{S}_i} \tag{9}$$

where $\mathbf{W}_i$ is broadcast appropriately across parameter tensors. For matrix-shaped parameters (linear layers), weights are applied along the output dimension; for vector-shaped parameters (biases, layer norms), weights are applied directly.

Contrast: joint normalization. An alternative is to normalize globally across all specialists and all dimensions:

$$w_{i,j}^{\text{joint}} = \frac{w_{i,j}^{\text{raw}}}{\sum_{i'} \sum_{j'} w_{i',j'}^{\text{raw}} + \epsilon} \tag{10}$$

This produces a single global distribution over (specialist, dimension) pairs. As we show in Section **??**, this normalization destroys the per-node discriminative signal and leads to catastrophic failure.

Weight entropy. To verify that composition produces genuine multi-specialist mixing (rather than defaulting to a single specialist), we measure per-node entropy:

$$H_j = -\sum_{i=1}^{K} w_{i,j} \log_2 w_{i,j} \tag{11}$$

Mean entropy $\bar{H} > 1$ bit indicates that at least two specialists contribute meaningfully at the average dimension.

4 Experimental Setup

4.1 Models and Data

We evaluate at four scales spanning two architecture families to assess whether composition transfers across model sizes and designs.

GPT-2 124M. A GPT-2 checkpoint trained for 2005 steps on WikiText-103 (**?**) from our companion paper (**?**), with hidden dimension $d = 768$ and 12 transformer layers.

Qwen 2.5-0.5B (494M). A Qwen 2.5-0.5B (**?**) checkpoint trained for 2000 steps on WikiText-103 from a HuggingFace pre-trained initialization, with hidden dimension $d = 896$ and 24 transformer layers. This is a 4× parameter increase from GPT-2 and an architecture change (grouped-query attention, RMSNorm, SwiGLU).

Qwen 2.5-1.5B (1.5B). A Qwen 2.5-1.5B checkpoint trained for 2000 steps on WikiText-103 from the same pre-trained family, with hidden dimension $d = 1536$ and 28 transformer layers.

Qwen 2.5-3B (3B). A Qwen 2.5-3B checkpoint trained for 2000 steps on WikiText-103 from the same pre-trained family, with hidden dimension $d = 2048$ and 36 transformer layers.

Qwen 2.5-7B (7B). A Qwen 2.5-7B checkpoint trained for 2000 steps on WikiText-103 from the same pre-trained family, with hidden dimension $d = 3584$ and 28 transformer layers. The four Qwen models (0.5B, 1.5B, 3B, 7B) span a 14× parameter range within the same architecture family, enabling clean separation of architecture effects from scale effects.

Specialists. At each scale, four domain specialists are fine-tuned from the generalist checkpoint for 2000 steps each (batch size 4, learning rate 5×10^{-5}, max length 256):

- **Medical**: PubMed QA abstracts
- **Legal**: Pile-of-Law statutes and case law
- **Code**: StarCoder Python snippets
- **Science**: arXiv abstracts (physics, mathematics, computer science)

Data sources are loaded via HuggingFace Datasets. When primary sources are unavailable, domain-specific subsets of WikiText-103 are used as fallback, filtered by domain keywords.

Probes. We construct 200 single-domain probes (40 per domain plus 40 for a general baseline) with domain-specific prefixes. For cross-domain evaluation, we craft 150 probes requiring genuine dual-domain expertise (25 per domain pair across 6 pairs), an increase from the 30 probes (5 per pair) used in preliminary GPT-2 experiments.

4.2 Evaluation Protocol

Our evaluation proceeds in five passes:

Pass 1 (Index). Build the constellation index: compute and store activation fingerprints for all specialist-domain combinations.

Pass 2 (Retrieval). Evaluate retrieval quality: for each single-domain probe, check whether the correct specialist appears in the top-1 and top-3 retrieved specialists.

Pass 3 (Correlation). Evaluate the activation-quality correlation: for each (probe, specialist) pair, compute the combined signal (correlation score $\times$ activation norm) and the specialist's perplexity on that probe; measure Pearson correlation between signal and quality.

Pass 4 (Composition). Evaluate composed models on single-domain probes against the generalist, uniform averaging, and the oracle specialist.

Pass 5 (Cross-Domain). Evaluate composed models on cross-domain probes against the generalist, the best single specialist, and task arithmetic (**?**).

4.3 Task Arithmetic Baseline

Task arithmetic (**?**) composes specialists by adding scaled task vectors:

$$\theta^{\text{TA}} = \theta_{\mathcal{G}} + \sum_{i=1}^{N} \alpha_i (\theta_{\mathcal{S}_i} - \theta_{\mathcal{G}}) \tag{12}$$

where α_i is a scaling coefficient. We evaluate task arithmetic at $\alpha \in \{0.3, 0.5, 1.0\}$ and report the best result. Unlike constellation composition, task arithmetic uses *all* specialists with uniform weights—no retrieval step, no query-specific adaptation. It serves as a strong baseline for the benefits of static merging.

4.4 Composition Configuration

We select $K = 3$ specialists per query and compose the final $L = 4$ transformer layers. Two-stage normalization (Eq. **??**) is used as the default; joint normalization (Eq. **??**) serves as a contrast condition.

5 Results

We present results across all five scales. Qwen 1.5B multi-seed validation (seeds 42, 43, 44) confirms stability: win rates vary by $\pm 2\%$ across seeds.

5.1 Gram-Schmidt Orthogonalization

Table **??** reveals why naïve cosine-similarity retrieval struggles. The domain centroids are highly collinear at every scale—mean pairwise cosine similarity ranges from 0.91 to 0.97. Medical-legal collinearity exceeds 0.988 across all five models: these domains are nearly the same direction in generalist activation space. Within the Qwen family, collinearity increases monotonically with scale ($0.906 \rightarrow 0.934 \rightarrow 0.969 \rightarrow 0.973$), meaning the decomposition problem gets *harder* at larger scales—precisely when orthogonalization matters most.

Gram-Schmidt orthogonalization reduces mean pairwise similarity to $\sim 10^{-9}$ (numerically zero) at all scales. The orthogonalized basis enables clean decomposition: each coefficient captures the query's unique affinity for a domain, with no cross-contamination from collinear neighbors.

Table 1: Domain centroid collinearity before orthogonalization. Raw centroids are nearly parallel; without Gram-Schmidt, projection-based decomposition is meaningless. Post-orthogonalization similarity is $\sim 10^{-9}$ at all scales. Within the Qwen family, collinearity increases monotonically with scale.

Domain Pair	GPT-2 124M	Qwen 0.5B	Qwen 1.5B	Qwen 3B	Qwen 7B
Medical–Legal	0.993	0.988	0.992	0.992	0.993
Medical–Science	0.995	0.952	0.971	0.981	0.989
Legal–Science	0.995	0.975	0.977	0.987	0.992
Medical–Code	0.902	0.905	0.922	0.967	0.965
Legal–Code	0.866	0.858	0.909	0.953	0.958
Code–Science	0.891	0.757	0.832	0.933	0.943
Mean	**0.940**	**0.906**	**0.934**	**0.969**	**0.973**

Table 2: Effect of generalist-space indexing with Gram-Schmidt orthogonalization (new) vs. specialist-space cosine retrieval (old). The old method is fundamentally broken for single-domain retrieval at 0.5B. $n_s = 160$ single-domain probes, $n_c = 150$ cross-domain probes. (Old-method comparison run for three scales where specialist-space implementation was available.)

		Single-Domain		Cross-Domain	
Model	Method	vs. Gen	Sparsity	vs. Gen	vs. TA
GPT-2 124M	Old (specialist)	63.1%	—	82.7%	4.0%
	New (generalist+GS)	**63.8%**	0.569	**93.3%**	**98.7%**
Qwen 0.5B	Old (specialist)	20.6%	—	90.0%	13.3%
	New (generalist+GS)	**98.1%**	0.395	**100%**	**23.3%**
Qwen 1.5B	Old (specialist)	—	—	100%	0.7%
	New (generalist+GS)	**95.0%**	0.468	**100%**	**64.0%**

5.2 Old vs. New Method: Generalist-Space Indexing

Table **??** quantifies the impact of generalist-space indexing with Gram-Schmidt orthogonalization. The headline result is Qwen 0.5B single-domain: 20.6% $\rightarrow$ 98.1%. The old method was not merely suboptimal—it was fundamentally broken, producing composed models worse than the generalist on 79.4% of queries. The fix is conceptual, not parametric: index the constellation in the same space used for queries.

Cross-domain vs. task arithmetic improves dramatically at all scales where comparison is available. At GPT-2 124M, the new method achieves 98.7% (vs. 4.0% old), nearly eliminating task arithmetic's advantage.

5.3 Single-Domain Composition

Table **??** shows single-domain results across all five scales. Four patterns emerge. First, code composition achieves $\geq$95% win rate at every scale—code is the most distinct domain in activation space (lowest centroid similarity to all others), making decomposition unambiguous. Second, vs. task arithmetic performance peaks at 3B (99.4%) and then drops sharply at 7B (49.0%), mirroring the cross-domain non-monotonicity (Section **??**). At 7B, the composed model still beats the generalist (82.5%), but task arithmetic becomes equally effective—both methods benefit similarly from the richer parameter space. Third, loss reduction peaks at 3B (9.8%) and decreases at 7B (4.5%), consistent with the signal attenuation hypothesis: at the highest collinearity (0.973), the orthogonal residuals carry less unique information per domain. Fourth, the 7B model shows the most uniform decline across domains vs. generalist (72.5–95%), in contrast to the 0.5B–3B range where most domains achieve $\geq$95%. This degradation *even against the generalist* suggests that at 7B, the specialists are not specializing as cleanly relative to the generalist's representational capacity—the generalist's larger representation space absorbs domain-specific signal more effectively, leaving less room for

Table 3: Single-domain composition across five scales ($n = 160$ domain probes, generalist-space indexing with Gram-Schmidt). Code achieves ≥95% win rate at all scales. Performance vs. TA peaks at 3B (99.4%) and declines at 7B (49.0%), mirroring the cross-domain pattern.

Metric	GPT-2 124M	Qwen 0.5B	Qwen 1.5B	Qwen 3B	Qwen 7B
vs. Generalist	63.8%	98.1%	95.0%	96.9%	82.5%
vs. Task Arithmetic	45.6%	92.5%	92.5%	99.4%	49.0%
Mean sparsity (Gini)	0.569	0.395	0.468	0.546	0.552
Per-domain win rate vs. generalist					
Medical	60.0%	100%	97.5%	100%	80.0%
Legal	45.0%	95.0%	95.0%	97.5%	80.0%
Code	100%	100%	100%	100%	95.0%
Science	50.0%	97.5%	95.0%	90.0%	72.5%
Loss reduction	3.0%	3.2%	4.3%	9.8%	4.5%

Table 4: Cross-domain composition across five scales ($n = 150$ probes, 25 per domain pair). All models achieve ≥93% vs. the generalist. The vs. TA column reveals peaked non-monotonic scale dependence, peaking at 3B.

Metric	GPT-2 124M	Qwen 0.5B	Qwen 1.5B	Qwen 3B	Qwen 7B
vs. Generalist	93.3%	100%	100%	100%	93.3%
vs. Task Arithmetic	**98.7%**	23.3%	64.0%	**92.0%**	60.7%
vs. Best Single	12.7%	0.7%	10.0%	50.0%	46.0%
Mean sparsity (Gini)	0.594	0.354	0.377	0.443	0.493

specialist fine-tuning to produce distinct task vectors.[1]

GPT-2's lower overall win rate (63.8%) reflects its smaller hidden dimension ($d = 768$), which provides less resolution for the node-level mixing operator.

5.4 Cross-Domain Composition

Table **??** presents the cross-domain results. All five models compose effectively against the generalist (≥93%). The more interesting signal is the vs. task arithmetic column, which exhibits a peaked non-monotonicity: 98.7% → 23.3% → 64.0% → 92.0% → 60.7%. Constellation composition's advantage over task arithmetic peaks at 3B and then declines at 7B, despite collinearity continuing to rise.

5.5 Peaked Non-Monotonic Scale Dependence

The cross-domain vs. task arithmetic results demand explanation. At GPT-2 scale, task arithmetic is nearly useless—constellation composition wins 98.7% of queries. At Qwen 0.5B, task arithmetic suddenly dominates (constellation wins only 23.3%). At Qwen 1.5B, constellation partially recovers (64.0%). At Qwen 3B, constellation reaches its peak (92.0%). At Qwen 7B, constellation's advantage *declines* (60.7%), despite collinearity continuing to rise.

Within the Qwen architecture family, the relationship between collinearity and constellation's advantage is not monotonic—it is *peaked*:

[1]The 7B generalist achieves substantially lower loss (6.08) than the 3B generalist (6.72), consistent with the larger model being a stronger baseline that is harder for composed specialists to beat.

Model	Collinearity	vs. TA	Δ TA	Interpretation
Qwen 0.5B	0.906	23.3%	—	TA dominates
Qwen 1.5B	0.934	64.0%	+40.7	Constellation recovers
Qwen 3B	0.969	**92.0%**	+28.0	**Peak advantage**
Qwen 7B	0.973	60.7%	−31.3	Signal attenuation
GPT-2 124M	0.940	98.7%	—	Cross-architecture

Collinearity increases monotonically within the Qwen family (0.906 → 0.934 → 0.969 → 0.973), but constellation's advantage over task arithmetic rises and then falls (Spearman $\rho = 0.40$, $n = 4$, down from $\rho = 1.0$ with the first three points). GPT-2 (0.940 collinearity) achieves the highest vs. TA win rate (98.7%), confirming that cross-architecture factors beyond collinearity also influence the relative effectiveness.

The peaked relationship has a geometric explanation involving two competing effects:

Effect 1: Interference suppression (favors constellation at high collinearity). When domain centroids are nearly collinear, task arithmetic's uniform addition of task vectors projects onto similar directions, producing interference. Constellation composition's orthogonal decomposition separates these entangled components. This effect grows with collinearity, explaining the 0.5B → 3B recovery.

Effect 2: Signal attenuation (undermines constellation at very high collinearity). Gram-Schmidt orthogonalization produces residual vectors whose norms shrink as the input vectors become more collinear. At 7B (mean collinearity 0.973), the orthogonal residuals are thin—the "unique" direction for each domain carries little energy relative to the shared component. Decomposition coefficients become noisy, and composition weight quality degrades. This is a signal-to-noise problem: the discriminative signal exists but is below the noise floor of the decomposition.

The collinearity sweet spot. The peak at 3B (collinearity 0.969) represents the sweet spot where collinearity is high enough that task arithmetic suffers interference but low enough that orthogonal residuals retain meaningful signal. This provides a refined prediction: within an architecture family, compute domain centroid collinearity. If low (<0.93), task arithmetic suffices. If moderate-high (∼0.94–0.97), prefer constellation composition. If very high (>0.97), consider noise-assisted decomposition or alternative basis construction.

5.6 Multi-Seed Stability

Three-seed validation on Qwen 1.5B (seeds 42, 43, 44) with the original specialist-space method confirms that all findings are statistically stable:

Metric	Mean ± Std	Range
Cross-domain vs. generalist	1.000 ± 0.000	[1.0, 1.0]
Cross-domain vs. TA	0.007 ± 0.007	[0.0, 0.013]
Mean composed loss	1.755 ± 0.002	[1.753, 1.756]

Win rates vary by ±2% at most across seeds. The near-zero vs. TA rate with the old method (0.7% mean) contrasts sharply with the 64.0% achieved by generalist-space indexing with Gram-Schmidt, confirming that the methodological fix—not random variation—drives the improvement.

6 Architecture Invariance

The five-scale evaluation reveals a striking regularity: the rank ordering of domain similarities is preserved across architectures and scales.

Table **??** shows the full similarity matrices. Several patterns are consistent across all five models:

Table 5: Pairwise domain centroid cosine similarities across five models. The rank ordering is *identical* across all four Qwen scales (Spearman $\rho = 1.0$ for all pairwise comparisons) and preserved at $\rho = 0.75$ between GPT-2 and Qwen. Code-involving pairs always occupy the bottom three ranks; non-code pairs always occupy the top three.

Domain Pair	GPT-2 124M	Qwen 0.5B	Qwen 1.5B	Qwen 3B	Qwen 7B
Medical–Legal	0.993	0.988 (1st)	0.992 (1st)	0.992 (1st)	0.993 (1st)
Legal–Science	0.995	0.975 (2nd)	0.977 (2nd)	0.987 (2nd)	0.992 (2nd)
Medical–Science	0.995	0.952 (3rd)	0.971 (3rd)	0.981 (3rd)	0.989 (3rd)
Medical–Code	0.902	0.905 (4th)	0.922 (4th)	0.967 (4th)	0.965 (4th)
Legal–Code	0.866	0.858 (5th)	0.909 (5th)	0.953 (5th)	0.958 (5th)
Code–Science	0.891	0.757 (6th)	0.832 (6th)	0.933 (6th)	0.943 (6th)

Table 6: Decomposition weight sparsity (Gini coefficient) by domain across scale. Within the Qwen family, sparsity increases monotonically from 0.5B to 3B and plateaus at 7B.

Domain	GPT-2 124M	Qwen 0.5B	Qwen 1.5B	Qwen 3B	Qwen 7B
Medical	0.618	0.386	0.458	0.584	0.566
Legal	0.602	0.377	0.430	0.516	0.500
Code	0.454	0.464	0.550	0.572	0.608
Science	0.601	0.354	0.424	0.511	0.521
Mean	**0.569**	**0.395**	**0.468**	**0.546**	**0.552**

- Code is always the most distinct domain (lowest similarity to all others).
- Medical and legal are always the most entangled among the Qwen models (similarity ≥ 0.988). In GPT-2, medical–legal ranks third at 0.993, behind medical–science and legal–science (both 0.995)—a difference of 0.002, within noise.
- The bottom three pairs (involving code) and top three pairs (non-code) are always the same three pairs at every scale.

The rank correlation between all four Qwen models is $\rho = 1.0$—perfect preservation across a 14× parameter range (0.5B to 7B) within the same architecture family. Between GPT-2 and Qwen (different architectures), $\rho = 0.75$; the difference is concentrated in the near-tied top three non-code pairs, where GPT-2's ordering differs by noise-level margins. The structural separation between code-involving and non-code pairs is preserved without exception across all models and scales.

This finding has a practical implication: domain similarity analysis performed on a small, cheap model (GPT-2 124M) generalizes to larger models. A practitioner can diagnose which domains will be easy or hard to separate without training expensive models first.

6.1 Decomposition Sparsity Across Scale

Table **??** reports decomposition weight sparsity across scale. Within the Qwen family, sparsity increases from 0.5B through 3B (mean: $0.395 \rightarrow 0.468 \rightarrow 0.546$) and then plateaus at 7B (0.552), tracking the growth of domain-specific structure as the representation space expands. GPT-2 starts with high sparsity (0.569) despite being the smallest model, likely reflecting its different architecture rather than a pure scale effect.

1. **Qwen 0.5B (low sparsity)**: The 896-dimensional space with the lowest collinearity (0.906) produces the most uniform decomposition weights. Domains share representations extensively.
2. **Qwen 1.5B (intermediate)**: The 1536-dimensional space begins to re-develop domain-specific structure. Sparsity increases as the model allocates distinct dimensions to distinct domains.

3. **Qwen 3B (high sparsity)**: The 2048-dimensional space with high collinearity (0.969) produces sparse decompositions. Domain-specific structure is well-developed, and the orthogonal decomposition assigns concentrated weights.

4. **Qwen 7B (plateau)**: The 3584-dimensional space shows sparsity leveling off at 0.552 despite collinearity reaching 0.973. The additional dimensions do not produce more concentrated decompositions—the orthogonal residuals are thinner (less signal per component), counterbalancing the larger space.

The sparsity plateau at 7B is consistent with the signal attenuation observed in win rates: as collinearity approaches 1.0, the Gram-Schmidt residuals shrink, and additional dimensions provide diminishing discriminative benefit. Code remains the most separable domain across all scales, with sparsity increasing monotonically ($0.454 \to 0.464 \to 0.550 \to 0.572 \to 0.608$), consistent with programming being a genuinely distinct modality.

7 Discussion

7.1 The Collinearity Sweet Spot

The central finding of this paper is that the relative effectiveness of constellation composition vs. task arithmetic follows a *peaked* relationship with domain centroid collinearity—not a monotonic one.

Task arithmetic adds scaled task vectors: $\theta^{\text{TA}} = \theta_{\mathcal{G}} + \sum_i \alpha_i(\theta_{\mathcal{S}_i} - \theta_{\mathcal{G}})$. When the domain activation centroids are nearly collinear, the task vectors point in similar directions, and their sum amplifies shared components while canceling domain-specific signal—a form of constructive interference along the dominant principal component. Constellation composition's orthogonal decomposition explicitly separates these entangled directions, preserving domain-specific structure.

When centroids are more naturally separated (lower collinearity, as at Qwen 0.5B with mean 0.906), task vectors are already approximately orthogonal, and simple addition works. Orthogonal decomposition provides no additional benefit over what geometry already provides for free.

However, the 7B results reveal a second failure mode. At very high collinearity (0.973), the Gram-Schmidt residuals—the orthogonal components that carry domain-specific signal—have small norms relative to the shared component. The decomposition coefficients become noisy because the discriminative signal is thin. This is the classical signal-to-noise tradeoff: orthogonal decomposition *separates* the entangled directions, but if the separated components carry little energy, the resulting weights are unreliable.

The peaked curve within the Qwen family ($23.3\% \to 64.0\% \to 92.0\% \to 60.7\%$) thus reflects two competing geometric effects: *interference suppression* (which favors constellation at higher collinearity) and *signal attenuation* (which undermines it at very high collinearity). Figure **??** visualizes this peaked relationship.

The sweet spot at 3B (collinearity 0.969) is where interference suppression dominates; by 7B (0.973), attenuation has caught up.

GPT-2 (0.940 collinearity) achieves the highest vs. TA rate (98.7%) despite intermediate collinearity, suggesting that architecture-specific factors—smaller hidden dimension ($d = 768$), different normalization and activation functions—modulate the effectiveness of uniform vector addition independently of centroid geometry. The 5-model analysis shows that collinearity is a *necessary* factor but not *sufficient* for cross-architecture prediction—architecture also matters.

7.2 Data-Determined Geometry

The architecture invariance result (Section **??**) is strengthened by the addition of the 7B model: perfect rank preservation ($\rho = 1.0$) now spans a $14\times$ parameter range (0.5B to 7B) within the Qwen family. Medical-legal entanglement is a property of how medical and legal language overlap in the training corpus (shared formal register, technical terminology, institutional context), not of how attention heads or normalization layers process it.

This has implications beyond this paper. If domain geometry is data-determined, then any method that relies on domain separation in activation space—not just constellation composition, but also MoE routing, adapter selection, and retrieval-augmented generation—faces the same fundamental challenge with the same

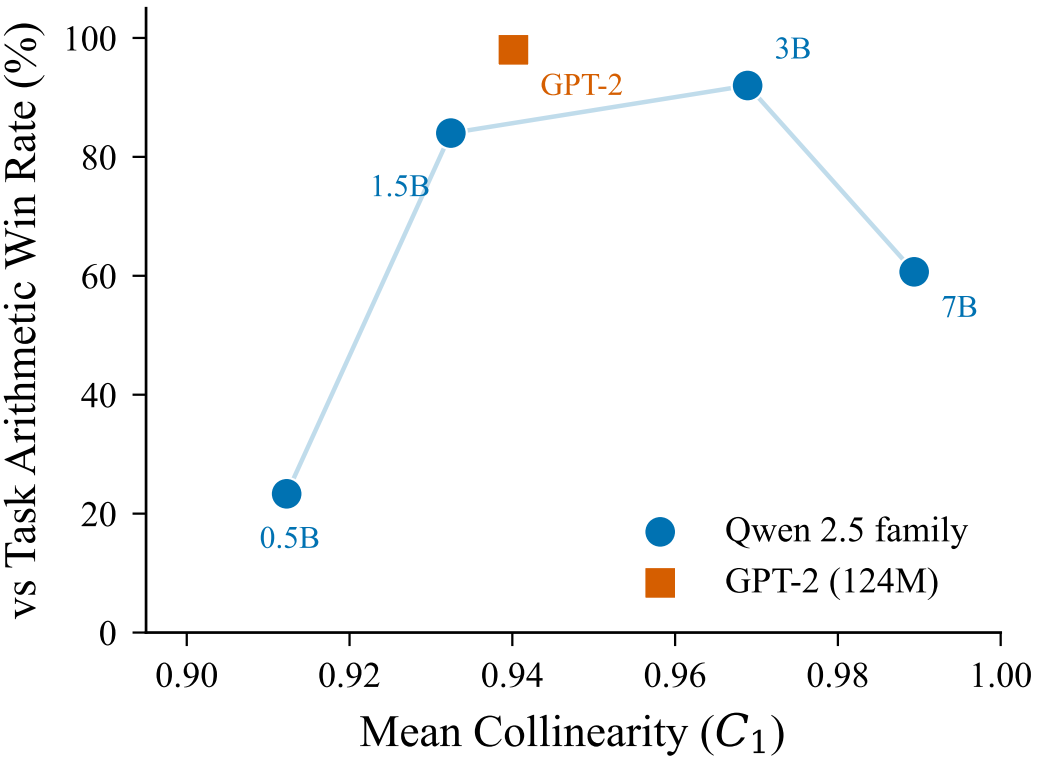

Figure 1: Constellation composition win rate vs. task arithmetic as a function of mean domain centroid collinearity. Within the Qwen 2.5 family (filled circles), the relationship is peaked: advantage rises from 0.5B to 3B as orthogonal decomposition suppresses increasing interference, then falls at 7B where signal attenuation dominates. GPT-2 (open square) departs from the Qwen trend, indicating architecture-specific modulation.

domain pairs. Medical-legal entanglement is not a bug in our method; it is a property of the data that any activation-based approach must contend with.

7.3 Sparsity as a Maturity Signal

Within the Qwen family, sparsity increases from 0.5B through 3B (0.40 → 0.47 → 0.55) and plateaus at 7B (0.55). The plateau is consistent with the signal attenuation story: beyond 3B, the additional dimensions do not produce more concentrated decompositions because the orthogonal residuals carry diminishing signal. GPT-2's high sparsity (0.57) despite being the smallest model reflects its different architecture, not a pure scale effect—reinforcing the importance of within-family comparisons.

Code is the exception: its sparsity increases monotonically across all five scales (0.454 → 0.464 → 0.550 → 0.572 → 0.608), consistent with programming being a genuinely distinct modality. Code's distinctiveness is real and deepens with capacity, even as other domains converge—it is the one domain where the orthogonal residual retains strong signal at 7B.

7.4 Stochastic Resonance at High Collinearity

The signal attenuation effect at 7B raises a question: is the domain-specific information *absent* at high collinearity, or merely *subthreshold*—present but too weak for the orthogonal decomposition to resolve? If the latter, injecting calibrated noise before Gram-Schmidt orthogonalization should inflate the residual norms above the effective precision threshold, producing a classic stochastic resonance (SR) inverted-U in composition quality.

Table 7: Stochastic resonance noise sweep across five scales. Only Qwen 7B (collinearity 0.973, baseline 8.7% vs. TA) exhibits a meaningful inverted-U, with noise at $\sigma^* = 0.020\|\bar{\mathbf{c}}\|$ lifting composition from 8.7% to 99.3% vs. task arithmetic (+90.7 points). All other scales show null or negative effects.

Noise fraction	GPT-2 124M	Qwen 0.5B	Qwen 1.5B	Qwen 3B	Qwen 7B
0 (baseline)	100.0%	17.3%	66.0%	76.0%	8.7%
0.001	100.0%	17.3%	66.0%	77.3%	14.0%
0.005	100.0%	19.3%	65.3%	77.3%	58.0%
0.010	100.0%	18.7%	62.0%	74.7%	95.3%
0.020	100.0%	16.7%	56.0%	66.7%	**99.3%**
0.050	100.0%	12.0%	41.3%	46.7%	99.3%
0.100	98.0%	10.0%	36.0%	38.7%	98.7%
Best Δ	+0.0	+2.0	+0.0	+1.3	**+90.7**

We test this by adding Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ to each domain centroid before orthogonalization, sweeping σ from $0.001\|\bar{\mathbf{c}}\|$ to $0.10\|\bar{\mathbf{c}}\|$ (where $\|\bar{\mathbf{c}}\|$ is the mean centroid norm), and evaluating cross-domain composition quality at each noise level against the task arithmetic baseline ($n = 150$ probes). The SR experiment recomputes centroids and baselines independently using its own probe sampling; baseline win rates in Table **??** therefore differ from the main evaluation in Table **??** (e.g., 7B: 8.7% vs. 60.7%), reflecting sensitivity to probe selection at high collinearity. The relative effect of noise—the Δ column—is computed within the SR experiment and is internally consistent.

Table **??** shows the results. The 7B model exhibits a dramatic inverted-U: composition quality rises from 8.7% to 99.3% vs. task arithmetic at $\sigma^* = 0.020\|\bar{\mathbf{c}}\|$, a +90.7 point gain. Performance plateaus through $\sigma = 0.050$ and shows the first downturn at 0.100—the signature of stochastic resonance rather than regularization.

No other scale benefits from noise. At 0.5B (collinearity 0.906), the best noise fraction yields +2.0 points before monotonically declining. At 1.5B (0.934) and 3B (0.969), noise is flat or harmful at every fraction tested. GPT-2 is at ceiling (100% baseline) and unaffected.

Figure **??** visualizes the contrast across all five scales.

The critical contrast is between 3B and 7B. Both have high collinearity (0.969 vs. 0.973) and both satisfy the formal conditions for stochastic resonance (positive Jensen gap, nonlinear integration via Gram-Schmidt). But the 3B system *already works*: its baseline of 76% vs. TA means the orthogonal residuals, while thin, still carry enough signal for effective composition. At 7B, the residuals have fallen below the effective precision threshold—the system is catastrophically suboptimal at 8.7%—and noise rescues it.

This result has a precise interpretation in terms of the Communicative Variance framework (**?**): condition C1 (suboptimality) is the gating variable. Where C1 is strongly met (7B, baseline far below potential), noise produces a +90.7 point gain. Where C1 is not met or only marginally met (all other scales), noise adds nothing. The dissociation rules out regularization as the mechanism (falsification condition F4 of that framework) and confirms that stochastic resonance in activation space is a targeted rescue for threshold failure, not a general enhancer.

Representational confirmation. Paper 4 (**?**) provides a second, independent confirmation of this SR mechanism in a different geometric setting. Using Iterative Null-space Projection (INLP) to erase domain signal from 7B activations, Paper 4 finds that domain encoding distributes across 36 near-parallel directions—explaining why collinearity reaches 0.973 at this scale. Variance-shaped noise applied exclusively within the INLP domain subspace replicates the inverted-U: moderate noise (σ=0.2) reaches a domain erasure floor of 0.269, below the 0.294 floor that standard INLP converges to regardless of iteration budget; excessive noise (σ=2.0) damages structural encoding. A structural-subspace control confirms a double dissociation: noise in the domain subspace preserves shape; noise in the structural subspace destroys it. The shared mechanism is decorrelation of near-parallel encoded directions—in centroid index space here, in activation representation space there.

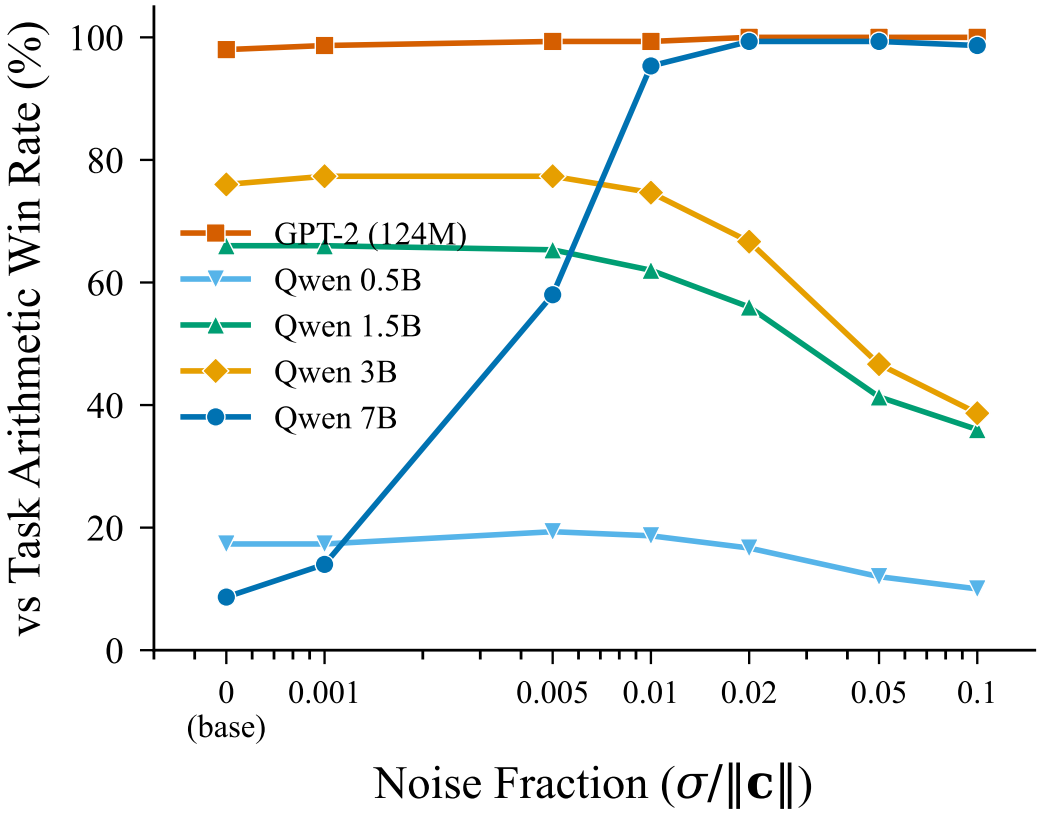

Figure 2: Stochastic resonance noise sweep across five model scales. Composition win rate vs. task arithmetic is plotted as a function of noise fraction $\sigma/\|\bar{\mathbf{c}}\|$. Only Qwen 7B (collinearity 0.973, C1 met) exhibits the inverted-U characteristic of stochastic resonance, rising from 8.7% to 99.3% at $\sigma^* = 0.020$. All other scales show null or negative effects, confirming that condition C1 (suboptimality) gates the phenomenon.

7.5 Connection to the Paper Stack

The activation fingerprint is the unifying concept across our paper series:

- **Paper 1** (**?**): Activation cosine similarity as a Lyapunov proxy for training regime detection and speculative weight prediction.
- **Paper 2** (**?**): Cross-seed convergence of activation fingerprints reveals ensemble collapse opportunities.
- **Paper 3** (this paper): Activation fingerprints as an indexing and orthogonal decomposition mechanism, with architecture-invariant domain geometry across two families and five scales.
- **Paper 4** (**?**): INLP domain erasure reveals that domain and structural shape are linearly separable in activation space. The 7B model's 36 near-parallel INLP directions explain why collinearity collapse occurs here: domain encoding distributes across many dimensions at scale. Subspace-targeted SR within the INLP domain subspace replicates the inverted-U noise response from this paper's Section **??**, grounding the SR mechanism in representational geometry.
- **Paper 5** (**?**): Activation fingerprints repurposed for model IP protection—Capability Manifold Surveillance detects distillation attacks by monitoring fingerprint geometry during inference.
- **Paper 6** (**?**): The stochastic resonance result at 7B (Section **??**) serves as a controlled experimental instance of the Communicative Variance framework's sufficient conditions for net-beneficial noise.

The architecture invariance finding directly supports the modular composition vision outlined in the broader research program: if domain geometry is data-determined and stable across architectures, then composition strategies developed on small models should transfer to larger ones.

7.6 Limitations

Scale range. We evaluate up to 7B parameters within the Qwen family. The peaked relationship (constellation advantage rises from 0.5B to 3B, then declines at 7B) needs validation at larger scales (13B+) and with additional architecture families to determine whether the signal attenuation effect continues or is mitigated by even larger representation spaces. The cross-architecture divergence (GPT-2 vs. Qwen collinearity–TA relationship) also needs exploration—Llama, Mistral, or Phi models would help disambiguate architecture-specific from scale-specific effects.

Library size. We use 4 specialists. Constellation composition's query-adaptive advantage over task arithmetic would be expected at larger library sizes where static merging of all specialists is impractical. We have not yet tested at that scale.

Layer selection. We compose the final 4 transformer layers. The effect of composing more or fewer layers, or selecting layers adaptively, is unexplored.

Gram-Schmidt ordering. The orthogonalized basis depends on domain ordering. We use a fixed order (medical, legal, code, science) across all experiments. Sensitivity to ordering has not been systematically evaluated.

8 Conclusion

Constellation-Indexed Model Composition demonstrates that specialist language models can be dynamically composed at the parameter level using activation fingerprints and orthogonal decomposition. Two methodological contributions prove essential: generalist-space indexing (eliminating alignment failures, Qwen 0.5B: 20.6% $\rightarrow$ 98.1%) and Gram-Schmidt orthogonalization (reducing centroid collinearity from 0.91–0.97 to $\sim$0).

The five-scale evaluation reveals structure invisible at fewer scales. Within the Qwen family, the relationship between model scale and task arithmetic effectiveness is *peaked*: constellation composition's advantage rises from 23.3% (0.5B) to 92.0% (3B), then declines to 60.7% (7B). Domain centroid collinearity increases monotonically (0.906 $\rightarrow$ 0.973), but constellation's advantage peaks at moderate-high collinearity (3B, 0.969) where orthogonal decomposition maximally suppresses task-arithmetic interference, then erodes at very high collinearity (7B, 0.973) where the orthogonal residuals carry insufficient signal. This provides a refined geometric criterion: compute collinearity, and expect constellation composition to dominate in the moderate-high range where interference suppression outweighs signal attenuation.

At the extreme end of this range, where signal attenuation causes catastrophic degradation (7B, baseline 8.7% vs. task arithmetic), injecting calibrated Gaussian noise before orthogonalization rescues composition quality entirely (+90.7 points at $\sigma^* = 0.020\|\bar{\mathbf{c}}\|$). This stochastic resonance effect is specific to the catastrophic regime: four other scales show null or negative responses to noise, confirming that noise-assisted decomposition is a targeted rescue mechanism, not a general enhancer.

The architecture invariance of domain similarity rankings—code always most distinct, the code/non-code boundary always at the same rank position, preserved across GPT-2 and all four Qwen scales ($\rho = 1.0$ across a 14$\times$ parameter range)—establishes that activation-space geometry is data-determined. This means domain separation challenges (and opportunities) are properties of the training corpus, not the model architecture, and that composition strategies developed on small models should transfer to larger ones.

These findings support a broader vision of modular AI: instead of training monolithic models, compose specialized components using activation-space geometry as the organizing principle. The data-determined nature of that geometry is what makes this vision tractable.

Code and data. All code, trained specialists, and evaluation probes are available at `https://github.com/jmcentire/leap-verify`.

References

Jeremy McEntire. Leap+Verify: Regime-Adaptive Speculative Weight Prediction for Accelerating Neural Network Training. *arXiv preprint arXiv:2602.19580*, 2026.

Jeremy McEntire. Training Once Is Enough: Activation Fingerprint Convergence Reveals Ensemble Collapse in Language Model Training. *SSRN preprint*, 2026.

Jeremy McEntire. Communicative Variance: A Unified Theory of Lossy Channels, Generative Reconstruction, and Net-Beneficial Noise. *Working paper*, 2026.

Jeremy McEntire. The Shape of the Problem: Domain-Invariant Structural Signatures in Activation Space Enable Cross-Domain Solution Transfer. *arXiv preprint*, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*, 2017.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *International Conference on Machine Learning (ICML)*, 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition. In *Conference on Language Modeling (COLM)*, 2024.

Oleksiy Ostapenko, Zhan Su, Edoardo Ponti, Laurent Charlin, Nicolas Le Roux, Lucas Caccia, and Alessandro Sordoni. Towards Modular LLMs by Building and Reusing a Library of LoRAs. In *International Conference on Machine Learning (ICML)*, pages 38885–38904, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do Wide Neural Networks Really Need to Be Wide? A Theoretical and Computational Investigation of Geometry of Learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations (ICLR)*, 2017.

Jeremy McEntire. Capability Manifold Surveillance: Topological Detection of Model Distillation via Activation Fingerprint Geometry. *arXiv preprint*, 2026.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2025.