Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Expert Collapse Is Not Inherited from Weights:\\Depth-Localized Re-Collapse across Three Mixture-of-Experts Architectures,\\and the Dimensional Artifact of the Weight--Activation Gap

□□ □

3078896787@qq.com

Wuxi Taihu University, School of Integrated Circuits, Department of Communication Engineering

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11744885/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Expert Collapse Is Not Inherited from Weights: Depth-Localized Re-Collapse across Three Mixture-of-Experts Architectures, and the Dimensional Artifact of the Weight–Activation Gap

Anonymous authors
Paper under double-blind review

Abstract

Kim (2026) reports a "weight–activation disconnect" in Mixture-of-Experts models: regularizing expert weights toward orthogonality leaves activation-space overlap unchanged, and activation MSO sits roughly 1,000× above weight MSO. The number is quoted as evidence that weight-space geometry carries no information about functional specialization.

We show that this ratio is an artifact of dimension. MSO is a mean of squared cosines, so for random vectors its expectation is $1/d$—and the two spaces being compared have $d \approx 10^6$–10^8 (expert weight matrices) versus $d \approx 10^2$–10^3 (residual-stream activations). Kim's 1,050× ratio is the same order as what two unrelated spaces of those dimensionalities produce (2,048× for his model). Normalized by the random floor, his model's ratio is 0.51×, and the activation space is in fact *less* aligned than the weight space.

Re-measuring five models spanning 130M to 47B parameters with a single metric, we find that the disconnect does not generalize. Floor-normalized, the ratio of activation to weight overlap spans six orders of magnitude across the five models, so it is a property of each model rather than of MoE. And three of them sit at the weight-space independence floor ($R_w = 1.1$–1.2), so their weight geometry carries no ranking at all, while their activation overlap spans a factor of 30: weight space stops discriminating before activation space does.

Our main result is depth-resolved, and it replicates across architectures. In all three 1B+ models we measure—Mixtral (8 experts, top-2), OLMoE (64 experts, top-8) and DeepSeek-MoE (64+2, top-6)—expert outputs are statistically independent through the middle of the network and then re-align sharply in the last one to three layers: swings of 16.7×, 336× and 65× against a flat weight profile. The magnitude of the final collapse varies by a factor of 22 across the three; its location does not. Over the same layers the weight geometry is either flat (Mixtral, 1.09× after consensus removal) or pinned at the floor (OLMoE, DeepSeek, $R_w \approx 1.2$). Collapse is generated in the forward pass near the output, not inherited from the weights of the layers in which it appears.

Finally, we show that the natural way to test this—subtracting the cross-expert consensus—has its own null distribution, $1/(E-1)^2$, which is 5×10^4 times larger than the independence floor when $E = 64$. Without it, centering manufactures overlap that is not there.

1 Introduction

Mixture-of-Experts (MoE) models scale capacity by activating only a subset of parameters per token Shazeer et al. (2017); Fedus et al. (2022); Lepikhin et al. (2021), and their value rests on an assumption that is rarely stated explicitly: that the experts specialize into functionally distinct sub-networks. A large body of work is built on this premise, whether it enforces diversity through routing (X-MoE Chi et al. (2022), HyperRouter Do et al. (2024), ReMoE Wang et al. (2025), CompeteSMoE Pham et al. (2024), SMoE-Dropout Chen et al.

(2023)), through optimization (orthogonal optimizers Liu et al. (2023), S2MoE Do et al. (2025)), or through geometric weight-space constraints Liu et al. (2023). Yet a growing set of measurements suggests the premise may not hold: expert representations can exhibit near-total similarity even in well-performing models Liu et al. (2023); Lo et al. (2025), and the degree of specialization "remains questionable" Lo et al. (2025).

Kim (2026) turns this skepticism into a controlled negative result. He regularizes expert weights toward mutual orthogonality across seven strengths. The intervention fails in two ways: weight-space overlap does not decrease (it rises by up to 114%), and activation-space overlap stays flat at an MSO of ≈ 0.57 regardless of regularization strength, with no significant correlation between the two ($r = -0.293$, $p = 0.523$). Kim concludes that nonlinearities break the connection between weight geometry and functional geometry, and proposes that the field should regularize in activation space instead.

That conclusion rests on a comparison between two numbers—activation MSO ≈ 0.57 and weight MSO $\approx 5 \times 10^{-4}$—read as a ratio of roughly 1,000×. In this paper we show that the ratio is not a property of the model. MSO is a mean of squared cosines between unit vectors, so its value under independence is $1/d$. Activation vectors live in $d = 512$ dimensions in Kim's model; flattened weight matrices live in $d = 1{,}048{,}576$. Two completely unrelated spaces of those sizes produce a ratio of $\approx 2{,}000\times$ by construction. When each space is normalized by its own floor, Kim's model shows a ratio of 0.51×: the weight space is the *more* aligned of the two.

This correction matters because the ratio has been read as a structural fact about MoE training. We therefore re-measure both spaces in five models spanning nearly three orders of magnitude—Kim's 130M model, Qwen1.5-MoE, Mixtral-8x7B, OLMoE and DeepSeek-MoE—using one metric and a floor-normalized statistic. Three results follow.

First, the disconnect does not reproduce. The raw ratio is the product of two factors—the ratio of ambient dimensions, $d_w/d_a \approx 10^3$–10^4, and the floor-normalized ratio R_a/R_w. It is the first factor, not the model, that sets the order of magnitude: Kim's 1,050× is 0.51× multiplied by his 2,048× dimension ratio. Normalized, the ratio spans six orders of magnitude across the five models (6×10^{-5} to 37), so "the gap between weight and activation overlap" is a property of individual models, not of MoE.

Second, weight space loses the ability to rank models once it reaches its floor. Three of the five models (Qwen, OLMoE, DeepSeek-MoE) have R_w between 1.12 and 1.24—statistically indistinguishable from independent random weights—while their activation overlap spans a factor of 30. Weight-space geometry is not merely a noisy proxy for functional specialization here; among these models it carries no signal at all. Kim's $r = -0.293$ is not in tension with this: it measures the effect of moving weights *within* one model through a regularizer, which is a different intervention from comparing models that differ everywhere at once.

Third, the collapse that remains is *depth-localized* in a way that its weights cannot explain, and this replicates. In all three 1B+ models, activation overlap is flat through the middle of the stack and then rises sharply over the final one to three layers: Mixtral from 0.051 to 0.856 (17×), OLMoE from 9.1×10^{-4} to 0.305 (336×), DeepSeek-MoE from 5.9×10^{-4} to 0.039 (65×). The magnitude of final collapse differs by 22× across the three; the position does not. Meanwhile the weight profiles are flat, and in the two 64-expert models they are flat at the independence floor. Whatever re-aligns the experts happens in the forward pass, at the end of the residual stream, and leaves no signature in the parameter geometry.

We also report a decomposition that shapes how the weight numbers should be read—and one caution about it. Expert weight matrices share a component across experts; for Mixtral, removing it takes the raw weight MSO of 0.2332 down to 0.0825. But the statistic that results has its own null distribution, $1/(E-1)^2$, an order of magnitude above the independence floor at $E = 8$ and five orders above it at $E = 64$: subtracting the mean forces the residuals to sum to zero, which manufactures apparent pairwise overlap of exactly that size. Read against that baseline, Mixtral's residual overlap is 4× real; OLMoE's and DeepSeek-MoE's are not distinguishable from the artifact at all.

1.1 Contributions

1. **We identify a dimensional artifact in the weight–activation gap.** Because MSO's null is $1/d$, ratios of activation MSO to weight MSO compare spaces whose baselines differ by three to

five orders of magnitude. Kim's $\approx 1{,}000\times$ becomes $0.51\times$ once normalized; the ordering of the two spaces reverses. The same statistic applied to our measurements gives $7.9\times$ for Qwen and 7×10^{-5} for Mixtral—spanning five orders of magnitude across three models, which is itself evidence that the ratio is not a stable property of MoE.

2. **We re-measure weight and activation overlap at 1B+ scale with one metric.** The raw activation-to-weight ratio factors into the dimension ratio d_w/d_a and the normalized ratio R_a/R_w; only the second is about the model. Across five models the normalized ratio spans six orders of magnitude. Three of the five sit at the weight-space floor ($R_w = 1.12$–1.24) while their activation overlap spans $30\times$: weight geometry stops discriminating before activation geometry does.

3. **Expert collapse is depth-localized while weight geometry is not, in three architectures.** Activation MSO rises in the last one to three layers of every model we measure—$17\times$ in Mixtral, $336\times$ in OLMoE, $65\times$ in DeepSeek-MoE—while weight MSO is flat over the same depth in all three ($1.09\times$ after consensus removal in Mixtral; at the floor throughout for the 64-expert models). The magnitude of the final collapse varies $22\times$; its position does not. Collapse is produced in the forward pass, not inherited from the weights of the layers where it appears.

4. **The consensus decomposition and its null.** Removing the cross-expert mean yields a statistic whose null is $1/(E-1)^2$—$10\times$ the independence floor at $E = 8$ and $5 \times 10^4\times$ at $E = 64$—because the residuals are constrained to sum to zero. Read against that baseline, Mixtral's residual weight overlap is $4\times$ real while OLMoE's and DeepSeek-MoE's are indistinguishable from the artifact. Cross-model comparisons of weight MSO must separate the two components *and* correct for this null.

2 Related Work

Expert collapse and routing diversity. Representation collapse in sparse MoE—the convergence of experts to redundant features—was named and analyzed by Chi et al. (2022), who show that routing-based training pulls token representations toward expert centroids. The mitigation literature is largely routing-side: auxiliary load balancing Fedus et al. (2022); Lepikhin et al. (2021), noise and capacity constraints, learned or hypernetwork routers Do et al. (2024), fully differentiable ReLU routing Wang et al. (2025), competitive assignment Pham et al. (2024), and fixed or progressively revealed routers Chen et al. (2023). Each of these changes which tokens reach which expert; none of them is a statement about the geometry of expert *outputs* for a common input, which is what our probe measures.

Weight-space diversity. A separate line regularizes the experts themselves: orthogonal expert optimizers that constrain update directions Liu et al. (2023), stochastic routing that perturbs the gate during training Do et al. (2025), and explicit orthogonality penalties Kim (2026). Kim (2026) is the closest antecedent to this paper and the one we engage with directly; our contribution relative to it is the normalization analysis and the depth resolution, not a new regularizer.

Measuring specialization. Lo et al. (2025) examine MoE internals and find that specialization is weaker than the routing literature assumes, with experts behaving as fine-grained neurons and routers favoring large-norm outputs. Our measurement follows the same commitment—specialization should be measured in the space where it would have to hold, namely the output of the experts—and adds the two things a single averaged number cannot provide: a floor to compare against and depth resolution.

Identifiability and alignment forces. The theoretical motivation for expecting collapse at all comes from identifiability results for joint-embedding objectives: under alignment with Gaussian regularization, learned representations are identifiable up to rotation, and the Gaussian is the unique distribution for which this holds Klindt et al. (2026). Cross-entropy on shared inputs has the same alignment structure, which makes a clone-like fixed point a natural expectation for expert outputs trained without any opposing pressure. We use this only as motivation for *where* to look—output space, and late in the network—and do not claim it as a tested mechanism here.

3 Method

3.1 Overlap and its null distribution

For N vectors, we use the mutual-subspace overlap of Kim (2026):

$$\mathrm{MSO}(\mathbf{A}) \;=\; \frac{2}{N(N-1)} \sum_{i<j} \big|\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle\big|^2, \tag{1}$$

where $\tilde{\mathbf{a}}_i$ is the L2-normalized flattening of entity i's parameter. We apply the identical form in both spaces: for weights, $\mathbf{a}_i$ is expert i's up-projection matrix $W_i^{\mathrm{up}} \in \mathbb{R}^{I \times C}$, flattened to $\mathbb{R}^{IC}$; for activations, it is expert i's output on a token, $\mathbb{R}^C$.

The statistic is bounded in $[0, 1]$ and reaches 1 at the clone state, which makes it tempting to compare values across spaces directly. That is not valid. For independent random unit vectors in $\mathbb{R}^d$, $\mathbb{E}[\cos^2] = 1/d$, hence

$$\mathbb{E}\big[\mathrm{MSO}\big] \;=\; 1/d \qquad \text{under independence.} \tag{2}$$

Weight matrices and residual-stream activations differ in d by three to five orders of magnitude, so their MSO values are not on a common scale. We therefore report the *normalized overlap*

$$R(\mathbf{A}) \;=\; d \cdot \mathrm{MSO}(\mathbf{A}), \tag{3}$$

which is ≈ 1 under independence and grows with genuine alignment, and we report raw MSO alongside it so that neither reading is hidden.

3.2 Consensus decomposition

Expert weight matrices in a trained MoE share a component across experts. Writing $W_i = \bar{W} + \Delta_i$ with $\bar{W} = \frac{1}{E}\sum_i W_i$, we compute MSO twice: once on $\{W_i\}$ (*raw*) and once on the residuals $\{\Delta_i\}$ (*consensus-removed*). The difference measures how much apparent expert similarity is a property of all experts jointly rather than of any expert's deviation from the group. The two parts are not additive in MSO—the statistic is quadratic—but reporting both separates a shared direction from expert-specific structure, which no single number does.

The centered statistic has its own null. It is tempting to compare the consensus-removed MSO against the same $1/d$ floor. That is wrong: centering changes the null. The residuals satisfy $\sum_i \Delta_i = 0$ identically, so they live in an $(E-1)$-dimensional subspace of the expert-index space regardless of d, and for independent noise

$$\mathbb{E}\big[\langle \tilde{\Delta}_i, \tilde{\Delta}_j \rangle\big] = -\frac{1}{E-1} \quad\Longrightarrow\quad \mathbb{E}\big[\mathrm{MSO}_{\mathrm{cent}}\big] \approx \frac{1}{(E-1)^2}, \tag{4}$$

independently of the ambient dimension. The offset is a negative cosine that squaring turns positive: centering *manufactures* apparent overlap of size $1/(E-1)^2$. For $E = 8$ that is 0.020, four orders of magnitude above $1/d$ for a flattened weight matrix; for $E = 64$ it is 2.5×10^{-4}, five orders above. A centered MSO is informative only if it exceeds this baseline, and all our reported values are read against it (we verified Eq. 4 on synthetic vectors: $E = 8$ gives 2.041×10^{-2}, $E = 64$ gives 2.57×10^{-4}).

3.3 Measurement procedure

Per-expert recomputation. We do not read off whatever the router happened to compute. For each MoE layer we capture the block input $\mathbf{h}$ and evaluate *all* N experts on *all* tokens, so every expert sees an identical input distribution. Pairwise differences are then attributable to expert parameters alone, with token-to-expert assignment removed as a confound; an MSO computed from routed outputs mixes expert similarity with routing-induced selection bias.

Two activation variants. Kim averages activation MSO over the experts the router co-activates for a token. We report both that quantity ($\mathrm{MSO}_{\text{top-}k}$, the Kim-comparable one) and the routing-agnostic $\mathrm{MSO}_{\mathrm{all}}$ over all $N(N-1)/2$ pairs.

Table 1: Model configurations. Total parameters are given, with activated parameters in parentheses where the two differ; Qwen1.5-MoE is a 14.3B-parameter model that activates 2.7B per token. The auxiliary coefficient is the value declared in each released configuration.

Model	**Params**	**Layers**	d_{model}	**Experts**	**top-k**	**Shared**	**aux.**
Kim (2026)	130M	6	512	8	2	no	0 (stated none)
Qwen1.5-MoE-A2.7B-Chat	14.3B (2.7B)	24	2048	60	4	yes	0.001
DeepSeek-MoE-16B	16.1B (2.8B)	28	2048	64	6	yes	—
OLMoE-1B-7B	6.9B (1.3B)	16	2048	64	8	no	—
Mixtral-8x7B	46.7B (12.9B)	32	4096	8	2	no	0.02

Sampling and precision. Activation statistics for the three 1B+ models use 8,192 tokens (32 chunks of 256) of held-out natural text (TruthfulQA questions, concatenated so that each chunk exceeds the 6–14 tokens of a single question); Qwen's were measured earlier at 2,048 tokens and are used only in Table 2. Models are loaded in bf16 with all parameters resident on one GPU and no quantization, so activation statistics carry no quantization error. Weight statistics are computed from the released checkpoints by streaming row blocks, exact in fp32 accumulation.

4 Experiments

We measure five models spanning nearly three orders of magnitude in parameter count and three routing regimes: the 130M baseline of Kim (2026); the three 1B+ models Qwen1.5-MoE, Mixtral-8x7B and OLMoE; and DeepSeek-MoE-16B, which adds a second shared-expert architecture. Mixtral matches Kim's routing geometry exactly—8 experts, top-2—while being 360× larger, which makes it the natural replication target. Table 1 lists the configurations.

4.1 The weight–activation gap is a dimension artifact

Table 2 gives the central comparison. The raw ratio of activation to weight MSO is the product of two factors,

$$\frac{\text{MSO}_{\text{act}}}{\text{MSO}_{\text{weight}}} = \underbrace{\frac{d_w}{d_a}}_{\text{dimension}} \cdot \underbrace{\frac{R_a}{R_w}}_{\text{normalized ratio}}, \tag{5}$$

and the dimension factor is 10^3–10^4 for every model in the table. Kim's 1,050× is 0.51× multiplied by his 2,048× dimension ratio; Qwen's 1,795× is 1.28× multiplied by 1,406×. The raw number therefore reports the ratio of the two ambient dimensions far more than it reports anything about the model.

Floor-normalizing both spaces (Eq. 3) removes that factor, and the picture inverts for Kim: his weight space is 569× above chance and his activation space 292× above chance, so the weights are the *more* aligned of the two. Across five models the normalized ratio ranges from 6×10^{-5} (Mixtral, where weight overlap is four orders of magnitude above its floor and activation is not) to 37 (OLMoE, the reverse)—six orders of magnitude. A quantity that varies over six orders of magnitude across five models is a per-model property, not a structural constant of MoE training.

4.2 Overlap tracks across spaces, and does not track scale

Read down the columns of Table 2 rather than across them, and the two spaces agree where they can and disagree where it matters. The two models whose weight overlap sits above the floor—Mixtral at 1.4×10^7 and Kim at 569—are also the two with the largest activation overlap (878 and 292). The three models pinned at the weight floor (Qwen 1.12, OLMoE 1.16, DeepSeek-MoE 1.24) cannot be ranked by weight geometry at all, yet their activation overlap spans a factor of 30 from 1.4 to 43. Weight space saturates: once expert weight matrices are statistically independent, the statistic has nothing left to say, while activation space continues to separate models that weight space calls identical.

Table 2: Weight- and activation-space overlap, raw and normalized by the independence floor $1/d$. $R = \text{MSO} \cdot d$ is ≈ 1 under independence; the last two columns decompose the raw ratio by Eq. 5. Activation columns are all-pairs statistics except for Kim, whose 0.570 is the top-2 statistic he reports (he gives no all-pairs value).

Model	d_w	**weight MSO**	R_w	**act MSO**	R_a	**raw ratio**	R_a/R_w
Kim 130M	1.0×10^6	5.43×10^{-4}	569	0.570	292	1,050×	0.51×
Qwen1.5-MoE	2.9×10^6	3.9×10^{-7}	1.12	7×10^{-4}	1.4	1,795×	1.28×
DeepSeek-MoE-16B	2.9×10^6	4.32×10^{-7}	1.24	2.4×10^{-3}	5.0	5,560×	4.0×
OLMoE-1B-7B	2.1×10^6	5.53×10^{-7}	1.16	0.0209	42.9	3.8×10^4×	37×
Mixtral-8x7B	5.9×10^7	0.2332	1.4×10^7	0.2144	878	0.92×	6.4×10^{-5}×

This is consistent with Kim (2026) rather than in conflict with it. His $r = -0.293$ is the effect of a within-model intervention that perturbs weights through a regularizer; such an intervention moves the weights without necessarily moving the outputs, because the output-space geometry is set by the interaction of all parameters with the data. Our comparison varies everything at once across models, and at the top of the range the two spaces move together. Both statements can be true, and together they say something sharper than either alone: weights and activations are coupled under training, decoupled under regularization, and weight space stops carrying information once it reaches its floor.

The ordering is not monotone in parameter count, in either normalization. Total parameters run 0.13B, 6.9B, 14.3B, 16.1B, 46.7B across the five models, while normalized activation overlap runs 292, 43, 1.3, 5.0, 878: the smallest model is second and the third-largest is last. Nothing supports the assumption—implicit on both sides of the debate—that specialization is a smooth function of scale. Since the five models differ simultaneously in routing geometry, expert count, training corpus and objective, we treat this as a natural experiment and draw no single-variable conclusion from it (§5).

4.3 Collapse is depth-localized; weight geometry is not

A per-model average hides the most striking structure in the data. Figure 1 resolves both spaces against normalized depth for the three 1B+ models, and Table 3 collects the summary statistics.

Activation space is strongly non-uniform, and in the same way three times. In Mixtral, overlap is near its minimum through the middle of the stack (0.0484 at layer 13) and then rises monotonically over the last five layers to 0.8368. In OLMoE it sits at 9.1×10^{-4}—indistinguishable from the 4.9×10^{-4} independence floor for 2048 dimensions—for thirteen consecutive layers, then reaches 0.3054 in the final layer, a jump that occurs essentially within that one layer. In DeepSeek-MoE the same profile appears with a smaller amplitude: 5.9×10^{-4} through the middle, 0.0386 in the last MoE block, with the transition concentrated in the final two blocks. The swings are 17×, 336× and 65×; the final-layer values differ by a factor of 22; the position does not vary at all.

Weight space behaves nothing like this. In Mixtral, raw weight MSO varies over a broad hump—0.365 near layer 10, 0.097 near layer 26, a 3.7× swing—but after the cross-expert consensus is removed the swing collapses to 1.09× (§4.4). In OLMoE and DeepSeek-MoE there is no hump to explain: their raw weight MSO is 5.5×10^{-7} and 4.3×10^{-7}, against independence floors of 4.8×10^{-7} and 3.5×10^{-7} ($R_w = 1.16$ and 1.24). Expert weight matrices in the two 64-expert models are, to measurement precision, independent random matrices at every depth—including the layers where the expert *outputs* become nearly identical. Whatever produces the terminal collapse is not a property of the parameters that carry it.

The residue is real but small. Removing the cross-expert consensus from the activations (§3.2) reduces the final-layer overlap from 0.856 to 0.148 in Mixtral, from 0.305 to 0.032 in OLMoE, and from 0.039 to 0.027 in DeepSeek-MoE. Against the centered null of $1/(E-1)^2$—0.020 for Mixtral's eight experts and 2.5×10^{-4} for the 64-expert models—those residuals are 7×, 130× and 110× real. Most of the terminal collapse is therefore a shared direction that all experts acquire in the last layers; a smaller but unambiguous expert-specific component appears alongside it.

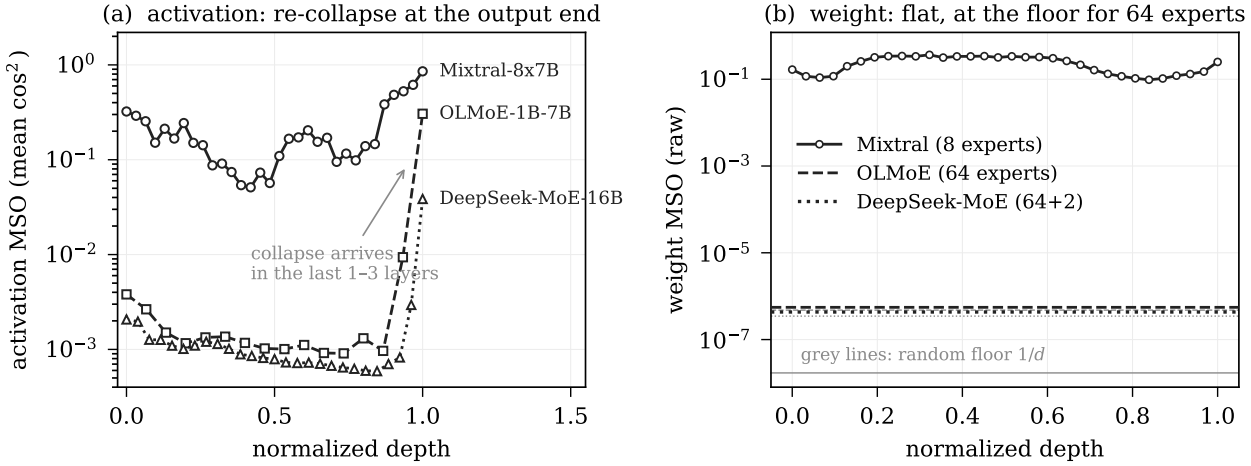


Figure 1: Overlap against normalized depth. (a) Activation-space MSO (mean $\cos^2$ over all expert pairs): in all three 1B+ models the experts are statistically independent through the middle of the stack and re-align sharply in the last one to three layers. (b) Weight-space MSO on the same depth axis, with each model's independence floor $1/d$ drawn as a thin grey line of the same style: Mixtral's weights vary smoothly over a hump, while the two 64-expert models sit at the floor at every depth. Note the log scale.

Table 3: Depth statistics. Activation: minimum over middle layers, final-layer value, and their ratio. Weight: raw level relative to the independence floor, and the depth swing of the raw profile. The magnitude of activated collapse varies 22× across models; its location and the flatness of the weight profile do not.

Model	**Layers**	**Experts**	**act min**	**act last**	**swing**	**weight** R_w
Mixtral-8x7B	32	8 / top-2	0.0512	0.8555	16.7×	1.4×10^7
OLMoE-1B-7B	16	64 / top-8	9.08×10^{-4}	0.3054	336×	1.16
DeepSeek-MoE-16B	27	64+2 / top-6	5.90×10^{-4}	0.0386	65.4×	1.24

4.4 Two-thirds of raw weight overlap is consensus—and the rest needs its own null

Removing the cross-expert mean reduces Mixtral's raw weight MSO from 0.2332 to 0.0825, a 65% reduction: roughly two-thirds of the raw number describes a direction that all eight experts share rather than any structure distinguishing them.

The residual must then be read against the centered null of Eq. (4), not against $1/d$, and the difference is decisive. Mixtral's 0.0825 sits 4.0× above its $1/(E-1)^2 = 0.020$ baseline, so a genuine expert-specific component survives—the raw number is mostly consensus, but not entirely. For the two 64-expert models nothing survives. Their centered weight MSO is 2.526×10^{-4} (OLMoE) and 2.530×10^{-4} (DeepSeek-MoE); the centered null is $1/63^2 = 2.520 \times 10^{-4}$; the ratios are 1.00 and 1.00. Two unrelated models, different parameter counts, different training data, the same number to three significant figures—which is the signature of a statistic reporting its own definition rather than the model.

The two components also behave differently with depth in Mixtral, and in a direction a "more similar weights, more similar outputs" account cannot use. Raw weight MSO peaks in the middle (layer 10) and falls to a minimum at layer 26, while activation MSO bottoms out in the middle (layer 13) and peaks at the end (layer 31). The correlation between the two profiles across the 32 layers is $r = -0.385$: if anything, the layers whose weights are most similar are the layers whose activations are least similar.

4.5 Robustness and sanity checks

Token budget. Mixtral's activation MSO was measured at three budgets. At 88 tokens (the raw question texts, 6–14 tokens each) the network mean is 0.2633 and layer 31 reaches 0.8417; at 2,048 tokens the mean

is 0.2145 and layer 31 reaches 0.8368; at 8,192 tokens the mean is 0.2144 and layer 31 reaches 0.8555, with the minimum at layer 12–13 in all three runs. The mean stabilizes once chunks are long enough to fill a sequence, and the depth profile—the quantity we interpret—is reproduced at every budget. The three 1B+ models use the 8,192-token run throughout.

Pairwise sanity. Point statistics could in principle be produced by a normalization bug, so we also report raw pairwise cosines. In activation space, Mixtral gives $\cos(\text{expert}_0, \text{expert}_1) = +0.616$ and $\cos(\text{expert}_0, \text{expert}_7) = +0.491$; Qwen gives $+0.007$ and -0.0001. In weight space, Mixtral's layer-0 up-projections give $\cos(\text{expert}_0, \text{expert}_1) = +0.189$ over the full 5.9×10^7-dimensional flattened matrix, computed by an independent code path. The two models differ in the raw statistic by the same orders of magnitude as in the aggregate.

Positive alignment, not noise. In activation space, Mixtral's mean signed cosine across expert pairs (0.4153) is essentially equal to its mean absolute cosine (0.4169): almost no pair points in opposing directions, which distinguishes partial collapse from chance correlation. Weight space shows the same magnitude of pairwise alignment (mean $|\cos| = 0.4673$); we did not compute the signed weight average.

What is not controlled. The five models differ in corpus, tokenizer, training duration, expert count, top-k, shared-expert structure and auxiliary objective, and none of these can be ablated on checkpoints we did not train. The comparison is a natural experiment. It establishes that certain claims are not general—that the gap ratio is not a model property, that weight space stops discriminating once it reaches its floor, and that neither collapse nor its absence is monotone in scale—but it cannot attribute any difference to a single variable. The depth result is the exception: within one checkpoint, width, corpus and tokenizer are fixed by construction, so the depth comparison is controlled even though the cross-model one is not.

5 Discussion

What the correction changes. The weight–activation ratio has been used as evidence that weight geometry is uninformative about functional specialization, and therefore that effort should move to activation-space regularization. The normalization analysis weakens the premise: the ratio is dominated by the difference in ambient dimension between the two spaces, and once that is removed Kim's own model no longer shows the effect. His intervention-level result—that regularizing weights does not move activation overlap—stands and is better explained by the intervention being upstream of the alignment force than by a dimensional argument. The distinction matters for what follows from it: a failed weight regularizer is evidence about that intervention, not about the relation between weights and activations in general.

Why the depth result is the strongest evidence here. Comparisons across models and across spaces both carry confounds: different dimensions, different training data, different routing. The depth comparison carries none of them. The layers of one checkpoint share a width, a floor, a corpus and a tokenizer; the only thing that varies is position. Under that control, activation overlap rises by 17× to 336× in the final layers of all three 1B+ models, while the weight profile over the same depth is flat—and in the two 64-expert models flat *at the independence floor*, meaning there is no weight structure in the final layers for the collapse to inherit. Whatever produces the late-layer re-collapse is a property of the forward computation at that depth, not of the parameters that carry it. The replication across three routing regimes is what raises this from an observation about one model to a regularity.

A candidate mechanism, held at arm's length. Cross-entropy on shared inputs applies an alignment force to expert outputs, and identifiability results for alignment objectives place the unopposed fixed point at a maximally correlated, Gaussian-like solution Klindt et al. (2026). In a residual network the output-near layers are where that force has both the strongest gradient and the shortest path to the loss, which is where we observe the re-collapse. This is a coherent account of why the effect is at the end rather than the beginning of the stack, and it is consistent with the middle layers being pushed away by routing and load-balancing pressure. It is not tested here: our three models confound that pressure with everything else, and a test requires training matched models with and without an opposing term.

Implications for measurement. Four practices follow. Report the independence floor alongside any MSO, because raw values from spaces of different dimension are not comparable. Separate consensus from expert-specific structure—and check the centered statistic against $1/(E-1)^2$, because below that baseline the decomposition measures the procedure. Resolve depth, because a single per-model average cannot distinguish "uniformly moderately similar" from "independent in the middle and identical at the end"—two situations with different causes and different remedies. And prefer activation space for any question about specialization: in three of our five models the weight statistic was at its floor and could not rank them, while the activation statistic separated them by a factor of 30.

Limitations. Weight statistics cover all layers for Mixtral, OLMoE and DeepSeek-MoE; for Qwen they were computed in the same session as its activation statistics, at sampled layers. Kim's model contributes published averages only, with no per-layer resolution available—precisely the resolution that shows the effect. The activation measurement for Qwen predates the unified probe used for the other three (it lacks the centered and top-k statistics), so it enters Table 2 but not Table 3. Our sample is five models, so cross-model orderings carry no statistical inference, and the natural-experiment caveat above applies throughout.

6 Conclusion

Expert collapse is real, but the statistic most often used to characterize it is not measuring what it appears to measure. Activation MSO and weight MSO live in spaces whose independence baselines differ by three to five orders of magnitude, and the "weight–activation gap" is largely that difference. The raw ratio factors into a dimension term and a floor-normalized term; only the second is about the model, and across five models it spans six orders of magnitude. There is no constant gap to explain.

What does survive, and what we take to be the paper's substantive finding, is depth. In all three 1B+ models we measure, experts are statistically independent through the middle of the stack and re-align sharply in the last one to three layers—by 17×, 336× and 65× in models with 8, 64 and 64+2 experts. The magnitude of the final collapse varies by a factor of 22; its position does not. Over the same layers the weight geometry is flat, and in the two 64-expert models it is flat at the independence floor: there is no weight structure in those layers for the collapse to inherit. A probe that reports one number per model cannot see any of this, and neither can a regularizer designed to act on the weights of the layers where the collapse is measured.

Two measurement cautions follow from the analysis, and both bit us during it. A statistic that subtracts a group mean acquires a null of $1/(E-1)^2$ rather than $1/d$, so centered overlap can be manufactured by the centering; and a weight statistic that saturates at its floor silently stops discriminating between models that differ by a factor of 30 in activation space. Both are cheap to check and both change the conclusion.

The obvious next step is the controlled version of this experiment: matched models trained with and without an opposing pressure, resolved by depth, in both spaces. Our data say where such an experiment should find its effect—the last few layers, in output space—and rule out at least the simplest alternative, that the experts become similar because their weights do.

References

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, & J. Dean (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR 2017*.

W. Fedus, B. Zoph, & N. Shazeer (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120), 1–39.

D. Lepikhin et al. (2021). GShard: Scaling giant models with conditional computation and automatic sharding. *ICLR 2021*.

Z. Chi et al. (2022). On the representation collapse of sparse mixture of experts. *NeurIPS 2022*.

G. Do, K. Le, Q. Pham, T. Nguyen, T.-N. Doan, B. T. Nguyen, C. Liu, S. Ramasamy, X. Li, & S. Hoi (2024). HyperRouter: Towards efficient training and inference of sparse mixture of experts. *arXiv:2312.07035*.

Z. Wang, J. Chen, & J. Zhu (2025). ReMoE: Fully differentiable mixture-of-experts with ReLU routing. *ICLR 2025*. *arXiv:2412.14711*.

Q. Pham, G. Do, H. Nguyen, T. Nguyen, C. Liu, M. Sartipi, B. T. Nguyen, S. Ramasamy, X. Li, S. Hoi, & N. Ho (2024). CompeteSMoE: Effective training of sparse mixture of experts via competition. *arXiv:2402.02526*.

T. Chen, Z. Zhang, A. K. Jaiswal, S. Liu, & Z. Wang (2023). Sparse MoE as the new dropout: Scaling dense and self-slimmable transformers. *ICLR 2023*. *arXiv:2303.01610*.

B. Liu, L. Ding, L. Shen, K. Peng, Y. Cao, D. Cheng, & D. Tao (2023). Diversifying the mixture-of-experts representation for language models with orthogonal optimizer. *arXiv:2310.09762*.

G. Do, H. Le, & T. Tran (2025). S2MoE: Robust sparse mixture of experts via stochastic learning. *arXiv:2503.23007*.

K. M. Lo, Z. Huang, Z. Qiu, Z. Wang, & J. Fu (2025). A closer look into mixture-of-experts in large language models. *Findings of ACL: NAACL 2025*, 4427–4447. *arXiv:2406.18219*.

H. Kim (2026). Geometric regularization in MoEs: The disconnect between weights and activations. *arXiv:2601.00457*.

D. Klindt, Y. LeCun, & R. Balestriero (2026). When Does LeJEPA Learn a World Model? *arXiv:2605.26379*. `https://arxiv.org/abs/2605.26379`