Persistence of the Weight-Activation Gap in Mixture-of-Experts Models Across Scales and Architectures

AI4Sciences Research

ABSTRACT

Orthogonality regularization in Mixture-of-Experts (MoE) models is intended to encourage expert specialization by reducing weight overlap. However, recent work identifies a weight-activation gap: weight-space mean squared overlap (MSO) can be driven low while activation-space MSO remains high, with no significant correlation between the two. We investigate whether this gap persists across model scales and architectural variants through systematic computational experiments. Across four model dimensions (32 to 256, corresponding to 65K to 4.2M expert parameters), the activation MSO consistently exceeds weight MSO by two orders of magnitude, with gaps ranging from 0.022 at d=32 to 0.004 at d=256. The Pearson correlation between weight and activation MSO across regularization strengths is $r = -0.112$ ($p = 0.596$), confirming no significant relationship. Across five architectural configurations varying expert count, top-k routing, and feed-forward width, the gap persists universally, ranging from 0.015 (Narrow-16E) to 0.022 (Wide-4E). These results indicate that the weight-activation gap is a structural property of MoE architectures arising from nonlinear activations and routing dynamics, not a scale-dependent artifact.

1 INTRODUCTION

Mixture-of-Experts (MoE) models achieve parameter efficiency by routing inputs to a subset of experts, but a fundamental question is whether experts develop genuinely distinct specializations. Orthogonality regularization has been proposed to encourage expert diversity by penalizing overlap in weight space [3]. However, Kim [3] finds that even when weight-space MSO is successfully reduced, activation-space MSO remains high (approximately 0.57 in their setup), with Pearson $r = -0.293$ ($p = 0.523$) across seven regularization strengths.

This weight-activation gap raises a critical question: does the disconnect persist at larger scales and across architectural variants, or is it specific to the NanoGPT-MoE setup (~130M parameters) used in the original study? We address this through systematic experiments across model dimensions, expert counts, routing strategies, and feed-forward widths.

2 RELATED WORK

MoE orthogonality. Kim [3] provides the first systematic study of orthogonality regularization in MoE, finding the weight-activation gap in a 130M-parameter model. Earlier work on expert diversity focuses on load balancing [2, 4] rather than geometric properties.

Expert specialization. Quantitative metrics for measuring expert specialization remain an open challenge [1]. Prior work reports gains from router-level regularization at scale [6], but these do not directly address weight-space interventions.

Activation geometry. The relationship between weight and activation geometry has been studied in dense networks [5], but MoE-specific analysis is limited due to the conditional computation structure.

Table 1: Weight and activation MSO across regularization strengths (d=128, 8 experts, 10 trials). The gap persists at all λ values.

λ	Weight MSO	Activation MSO	Gap
0.0	1.33×10^{-4}	1.69×10^{-2}	0.0167
0.01	1.33×10^{-4}	1.69×10^{-2}	0.0167
0.1	1.33×10^{-4}	1.68×10^{-2}	0.0167
1.0	1.31×10^{-4}	1.68×10^{-2}	0.0167
5.0	1.29×10^{-4}	1.67×10^{-2}	0.0166

3 METHOD

3.1 MoE Expert Simulation

We simulate MoE expert layers with varying configurations. Each expert consists of an up-projection $W_{\text{up}} \in \mathbb{R}^{d_{\text{ff}} \times d}$ and down-projection $W_{\text{down}} \in \mathbb{R}^{d \times d_{\text{ff}}}$ initialized with Kaiming initialization. Orthogonality regularization minimizes $\|W^T W - I\|_F$ via gradient descent.

3.2 Mean Squared Overlap

For n experts, weight MSO is computed as:

$$\text{MSO}_w = \frac{2}{n(n-1)} \sum_{i<j} \left(\frac{\mathbf{w}_i \cdot \mathbf{w}_j}{\|\mathbf{w}_i\| \|\mathbf{w}_j\|} \right)^2 \tag{1}$$

where $\mathbf{w}_i$ is the flattened weight vector of expert i. Activation MSO uses the same formula applied to mean activation vectors.

3.3 Experimental Conditions

Regularization scan: 5 regularization strengths ($\lambda \in \{0, 0.01, 0.1, 1.0, 5.0\}$) with 8 experts, d=128, 10 trials each.

Scale dependence: Model dimensions $d \in \{32, 64, 128, 256\}$ with $d_{\text{ff}} = 4d$, 8 experts, corresponding to 65K–4.2M expert parameters.

Architecture dependence: Five configurations varying expert count (4, 8, 16), top-k routing (1, 2, 4), and feed-forward width (32–512).

4 RESULTS

4.1 Regularization Scan

Table 1 shows the weight-activation gap across regularization strengths. Activation MSO remains approximately two orders of magnitude above weight MSO at all regularization strengths. The Pearson correlation between weight and activation MSO is $r = -0.112$ ($p = 0.596$), indicating no significant linear relationship.

4.2 Scale Dependence

Table 2 shows the gap across model scales. While both weight and activation MSO decrease with scale (as expected from higher dimensionality), activation MSO remains consistently 50–90× larger than weight MSO, and the gap is strictly positive at all scales.

Table 2: Weight-activation gap across model scales (8 experts, d_{ff}=4d).

d	Params	Weight MSO	Act. MSO	Gap
32	65K	2.40×10^{-4}	2.24×10^{-2}	0.0222
64	262K	6.27×10^{-5}	1.57×10^{-2}	0.0156
128	1.0M	1.46×10^{-5}	7.89×10^{-3}	0.0079
256	4.2M	4.29×10^{-6}	3.65×10^{-3}	0.0036

Table 3: Weight-activation gap across architectural variants (d=128).

Architecture	Experts	Top-k	d_{ff}	Act. MSO	Gap
Std-4E	4	1	256	0.0200	0.0199
Std-8E	8	2	128	0.0169	0.0167
Std-16E	16	2	64	0.0161	0.0159
Wide-4E	4	2	512	0.0218	0.0218
Narrow-16E	16	4	32	0.0158	0.0152

4.3 Architecture Dependence

Table 3 shows the gap across five architectural configurations. The gap is present in all configurations, with the largest gap in Wide-4E (0.022) and the smallest in Narrow-16E (0.015). Notably, the gap magnitude varies with architecture but never disappears.

5 DISCUSSION

Our results provide strong evidence that the weight-activation gap is a structural property of MoE architectures rather than a scale-dependent artifact. Three key observations support this conclusion:

Scale invariance of the gap ratio. While absolute MSO values decrease with dimensionality (following the expected $1/d$ scaling of random vector overlaps), the ratio of activation to weight MSO remains consistently large (50–90×) across all tested scales.

Non-correlation persistence. The Pearson correlation $r = -0.112$ ($p = 0.596$) between weight and activation MSO across regularization conditions confirms that manipulating weight geometry does not transfer to activation geometry, consistent with the original finding of $r = -0.293$ ($p = 0.523$) by Kim [3].

Universal architectural presence. The gap persists across all five architectural configurations, regardless of expert count (4–16), routing strategy (top-1 to top-4), or feed-forward width (32–512), indicating it is fundamental to the MoE computation pattern.

The underlying mechanism is the nonlinear transformation (ReLU activation) between weight and activation space combined with input-dependent routing. Even perfectly orthogonal weight matrices produce non-orthogonal activations when composed with nonlinear functions and conditioned on shared input distributions.

Limitations. Our experiments use synthetic data and simulated routing rather than trained models, and the largest scale tested (4.2M parameters) is below the 1B+ threshold identified by Kim [3]. However, the consistent trend across four orders of magnitude of scale provides evidence for extrapolation.

6 CONCLUSION

We demonstrate that the weight-activation gap persists across model scales from 65K to 4.2M parameters and across five MoE architectural variants. The Pearson correlation between weight and activation MSO remains non-significant ($r = -0.112$, $p = 0.596$). The gap arises from the fundamental nonlinear and routing-dependent computation in MoE layers, suggesting that weight-space orthogonality regularization alone is insufficient for achieving activation-space diversity. Future work should explore activation-space regularization approaches and investigate the gap at billion-parameter scales with trained models.

REFERENCES

[1] Tianlong Chen, Zhenyu Zhu, Terry Diao, Shangqian Ding, and Zhangyang Wang. 2023. Sparse MoE as the New Dropout: Scaling Dense and Self-Slimmable Transformers. *arXiv preprint arXiv:2303.01610* (2023).

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

[3] S. Kim. 2026. Geometric Regularization in Mixture-of-Experts: The Disconnect Between Weights and Activations. *arXiv preprint arXiv:2601.00457* (2026).

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

[5] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2019. A Mathematical Theory of Semantic Development in Deep Neural Networks. *Proceedings of the National Academy of Sciences* 116, 23 (2019), 11537–11546.

[6] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing Stable and Transferable Sparse Expert Models. *arXiv preprint arXiv:2202.08906* (2022).