MoE Align: Quantifying Expert Similarity Across Transformer Layers

Sai Vamsi Alisetti
Department of Computer Science
University of California, Santa Barbara
saivamsi@ucsb.edu

Abhishek Kumar
Department of Computer Science
University of California, Santa Barbara
abhishek_kumar@ucsb.edu

Abstract

Modern Mixture-of-Experts (MoE) Transformers allocate expert functions independently at each layer, assuming layer-wise specialization. This work investigates whether experts across different layers perform similar non-linear transformations, enabling potential expert reuse. We introduce an adapter-based method to align the input distribution between layers and evaluate functional similarity by training a lightweight adapter such that one layer's expert can approximate another's output. Empirical results on language modeling tasks reveal that certain cross-layer expert pairs exhibit high functional alignment, suggesting redundancy in depth. This finding opens avenues for parameter-efficient, depth-shared MoE architectures. Our code can be found here - GitHub.

1 INTRODUCTION

Mixture-of-Experts (MoE) models scale transformers by sparsely activating a subset of experts per token, significantly increasing parameter capacity without incurring additional inference cost. However, MoE layers are typically treated as depth-specific modules, with each layer containing an independently routed set of experts.

This architectural design assumes that each layer performs a distinct transformation. However, inspired by the Universal Transformer [3] and the inductive behavior of adjacent layers [5, 2], we explore the hypothesis that experts in different layers may be functionally equivalent.

To test this, we propose a simple but effective evaluation: using a small adapter to align the input space of one expert to that of another, and measuring output similarity through mean squared error (MSE). If a low-loss mapping exists, it implies that both experts are performing similar non-linear transformations, up to a change of input domain.

Our empirical study reveals that many expert pairs especially those in nearby layers exhibit strong alignment. These results suggest a surprising degree of depth redundancy and open the door to more parameter-efficient MoE models through expert reuse across layers.

2 RELATED WORK

Expert Similarity and Redundancy. Several studies have investigated expert redundancy within MoE models. [4] analyzed intra-layer expert similarity and proposed clustering-based reductions. [6] introduced mechanisms for early exiting and expert affinity scoring, though their focus was primarily on inference efficiency rather than functional equivalence.

Universal and Recurrent Transformers. The Universal Transformer [3] proposed reusing a single layer recurrently with shared weights, enriched by learned depth embeddings. Recent works like

MoEUT [1] combined this approach with MoE, showing that stacked transformer layers may be unnecessary and that learned recurrence can maintain competitive performance.

Inter-layer Functional Decomposition. Studies such as [5] and [2] highlight that transformer layers often collaborate to implement complex reasoning steps. These findings provide a foundation for our hypothesis: if adjacent layers are complementary, their experts may be interchangeable up to a transformation of the input distribution.

3 PROBLEM FORMULATION

Let $f_{\ell,e} : \mathbb{R}^d \to \mathbb{R}^d$ denote the non-linear transformation implemented by expert e in layer ℓ of a Mixture-of-Experts (MoE) transformer, where d is the hidden dimension. Token representations routed to different experts across layers are drawn from distinct distributions due to intervening attention and gating operations. Let $\mathcal{D}_{\ell,e}$ denote the empirical input distribution of expert (ℓ, e).

We investigate whether two experts f_{ℓ_1,e_1} and f_{ℓ_2,e_2} residing at different layers can be functionally aligned under an input transformation. Specifically, we ask whether there exists a transformation $A : \mathbb{R}^d \to \mathbb{R}^d$ such that:

$$f_{\ell_2,e_2}(A(x)) \approx f_{\ell_1,e_1}(x), \quad \forall x \sim \mathcal{D}_{\ell_1,e_1}$$

This expresses the notion of *local functional equivalence under input reparameterization*: while f_{ℓ_1,e_1} and f_{ℓ_2,e_2} may differ globally, they may behave similarly over their respective input domains if $A(x) \sim \mathcal{D}_{\ell_2,e_2}$.

We instantiate A as a lightweight adapter consisting of a linear projection followed by LayerNorm:

$$A(x) = \mathrm{LayerNorm}(Wx + b)$$

with $W \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. This design allows expressive yet constrained alignment of distributions without degenerate warping. The adapter is trained to minimize the mean squared error (MSE) between expert outputs:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x\sim\mathcal{D}_{\ell_1,e_1}} \left[\| f_{\ell_2,e_2}(A(x)) - f_{\ell_1,e_1}(x) \|_2^2 \right]$$

A low reconstruction loss suggests that the composite function $f_{\ell_2,e_2} \circ A$ approximates f_{ℓ_1,e_1} on $\mathcal{D}_{\ell_1,e_1}$. This indicates that the experts are functionally similar in the input region they operate over, enabling potential reuse across layers.

To prevent trivial mappings, future extensions can incorporate distributional regularization e.g., KL divergence between $A(x)$ and $\mathcal{D}_{\ell_2,e_2}$, ensuring the adapter stays faithful to the target expert's domain.

This formulation provides a principled way to test inter-layer expert similarity in deep MoE architectures under the realistic constraint of distributional shift.

4 PRELIMINARY RESULTS

4.1 Experimental Setup

We conduct our experiments using the `Qwen1.5-MoE-A2.7B` model, a decoder-only mixture-of-experts (MoE) language model publicly available via the HuggingFace Hub. The model employs sparse MoE routing in the feedforward layers with top-k expert selection and consists of 2.7 billion parameters. All experiments are performed in `float16` precision on a single NVIDIA A100 GPU using HuggingFace's `transformers` library.

For data, we use a subset of English Wikipedia sentences, tokenized with the corresponding `QwenTokenizer`. The dataset is split 80/20 for training and validation of adapters.

We focus on comparing functional similarity between experts across two different layers. We denote an expert as $f_{\ell,e}$, where ℓ is the layer index and e is the expert ID. For each target expert pair (ℓ_1, e_1) and (ℓ_2, e_2), we extract:

- z_1: input to expert f_{ℓ_1,e_1} before routing,

- $f_{\ell_1,e_1}(z_1)$: expert output (ground-truth target),
- $f_{\ell_2,e_2}(A(z_1))$: output of expert e_2 at layer ℓ_2 when fed the aligned input via a learned adapter A.

Adapters are implemented as a single linear projection followed by layer normalization. These are trained to minimize mean squared error (MSE) between $f_{\ell_2,e_2}(A(z_1))$ and $f_{\ell_1,e_1}(z_1)$.

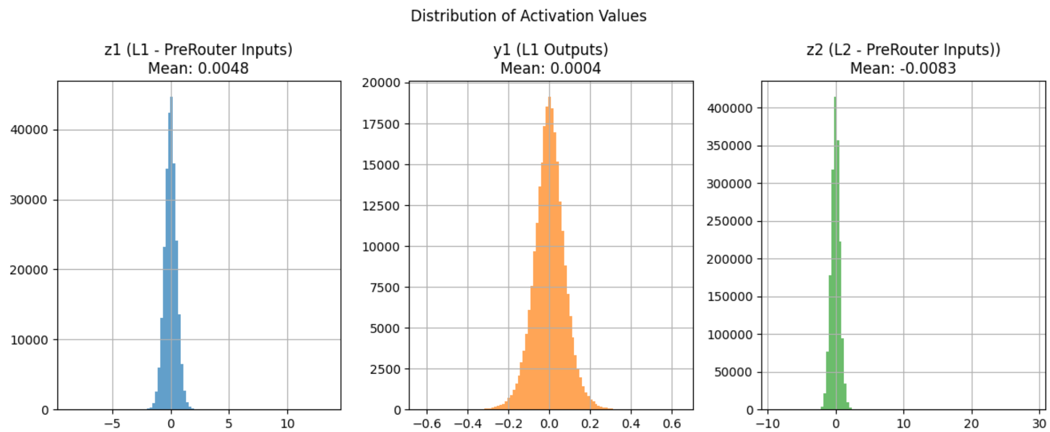

Figure 1: Distribution of Activation Values Across Layers. This figure shows histograms of token activation values from different stages in a Mixture-of-Experts (MoE) transformer. The left panel (z_1) displays the distribution of inputs to Layer 1 experts before the router. The middle panel (y_1) shows the output of Layer 1 experts. The right panel (z_2) shows the input to Layer 2 experts before routing. The distributions are approximately centered at zero but exhibit differing variances, indicating that each expert layer receives differently scaled input distributions. This highlights the need for learnable input transformations when comparing functional similarity between experts across layers.

4.2 Adapter-Based Functional Alignment

To evaluate functional similarity across layers, we learn an input adapter A such that $f_{\ell_2,e_2}(A(x)) \approx f_{\ell_1,e_1}(x)$ over inputs $x \sim \mathcal{D}_{\ell_1,e_1}$. We report the mean squared error (MSE) between expert outputs under this transformation.

Table 1 highlights a noteworthy finding: Layer 1 Expert 0 consistently exhibits high alignment loss (> 7.5) across all pairings with Layer 2 experts. In contrast, other Layer 1 experts (e.g., Experts 4, 8, 12) achieve alignment losses below 0.02 when mapped to suitable Layer 2 counterparts. This strongly suggests that Expert 0 at Layer 1 performs a functionally distinct computation that cannot be approximated via a simple adapter, indicating outlier behavior or specialized routing.

L1→L2	**E0**	**E4**	**E8**	**E12**	**E16**	**E20**	**E24**	**E28**
E0	8.34	7.71	7.81	7.87	7.88	7.83	7.76	7.62
E4	0.017	0.010	0.018	0.014	0.013	0.011	0.016	0.014
E8	0.017	0.011	0.016	0.014	0.013	0.012	0.014	0.012
E12	0.015	0.009	0.014	0.011	0.011	0.010	0.013	0.011

Table 1: MSE Losses between Layer 1 and Layer 2 expert pairs using input adapters. Expert **L1E0** exhibits significantly higher losses across all L2 experts, indicating it performs a distinct function relative to the rest.

5 Conclusion & Future Work

Our experiments reveal that while many experts across different layers of a Mixture-of-Experts (MoE) transformer exhibit functional redundancy, certain experts—such as Layer 1 Expert 0 (L1E0) demonstrate significantly higher adapter alignment loss, indicating a distinct computational role. This suggests that not all experts are functionally interchangeable, even under distribution-matching transformations.

We introduced a lightweight adapter mechanism to align expert input distributions and showed that low MSE alignment loss is a useful proxy for functional similarity.

In future work, we plan to:

- Extend swap-and-adapt experiments to all layers and include evaluation metrics such as perplexity degradation post-swap.
- Investigate fine-grained routing flexibility, where routers leverage functional similarity scores to dynamically choose experts from different layers.
- Explore sparsity and compression strategies informed by aligned expert clusters to reduce model redundancy.
- Examine the generalization behavior of adapted experts on out-of-distribution data to test robustness of functional equivalence.

Our results provide preliminary but compelling evidence that inter-layer expert reuse is feasible, and that functional specialization is unevenly distributed across layers. This opens up opportunities for more efficient and interpretable MoE transformer designs.

6 Appendix

While the main report analyses functional equivalence of experts through adapter-based interventions, we additionally ran a coarser stress test: swapping the entire sparse-FFN block (router + experts) between adjacent Transformer layers of a trained MoE. If two neighboring MoE blocks are functionally interchangeable, exchanging them should not harm task performance; a marked drop, on the other hand, suggests layer-specific specialization.

Experimental setup. We fine-tuned `google/switch-base-8` for three epochs on the GLUE MNLI training set, which is a multi-class classification task. We swapped all pairs of entire sparse-FFN blocks in the encoder portion of Switch-Base and recorded validation accuracies for each of them in Table 2. These results support the hypothesis that adjacent layers are functionally similar while layers farther apart are increasingly dissimilar.

Figure 2 plots the frequency of expert activation in all decoder layers for the baseline model and for three FFN swap variants. Exchanging the Layer-1 block with an adjacent Layer-3 block (1 <-> 3) leaves the pattern almost unchanged, whereas swaps with deeper layers (1 <-> 7, 1 <-> 9) progressively distort the routing distribution. The effect supports the view that the functional similarity between MoE blocks decays rapidly with layer distance.

Table 2: MNLI dev accuracy after swapping the sparse-FFN blocks of each adjacent layer-pair in `Switch-Base-8`. Bold diagonal = no swap.

	Layer 1	**Layer 3**	**Layer 5**	**Layer 7**	**Layer 9**	**Layer 11**
Layer 1	**0.77**	0.74	0.728	0.654	0.378	0.406
Layer 3		**0.77**	0.772	0.738	0.654	0.566
Layer 5			**0.77**	0.750	0.744	0.720
Layer 7				**0.77**	0.760	0.748
Layer 9					**0.77**	0.748
Layer 11						**0.77**

References

[1] Róbert Csordás, Zihan He, and Jürgen Schmidhuber. Moe-ut: Mixture-of-experts in a universal transformer. *arXiv preprint arXiv:2408.06793*, 2024.

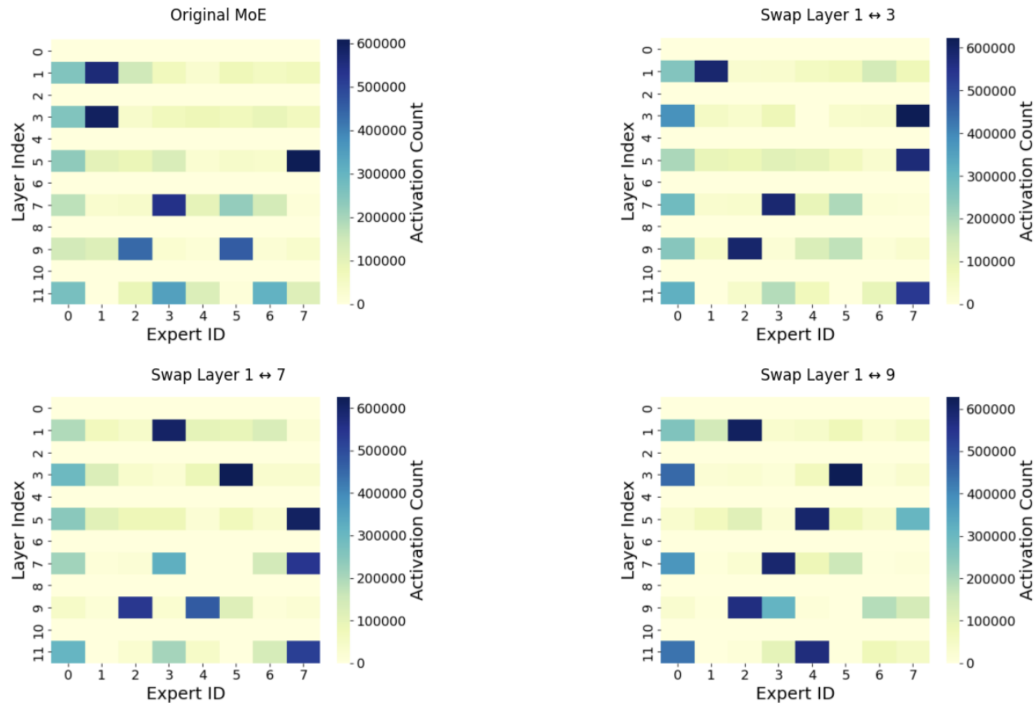

Figure 2: Token-routing heat-maps for the original model and after swapping Layer 1 with Layers 3, 7, 9. Darker cells denote higher activation counts.

[2] Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *International Conference on Learning Representations (ICLR)*, 2021.

[3] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Shuai Li, Xuezhi Wang, William Fedus, and et al. Are experts really needed? sampling-based inference for mixture-of-experts. *arXiv preprint arXiv:2310.01334*, 2023.

[5] Cliff Olsson, Deep Ganguli, Neel Nanda, and et al. In-context learning and induction heads. `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/`, 2022. Transformer Circuits Thread.

[6] Zeyuan Yao, Zihan He, Yan Xu, and et al. Deepmoe: Efficient and interpretable mixture-of-experts for llm inference. *arXiv preprint arXiv:2405.16039*, 2024.