ON THE SPATIAL STRUCTURE OF MIXTURE-OF-EXPERTS IN TRANSFORMERS

Daniel Bershatsky
AI Center, Skoltech
Moscow, Russia
daniel.bershatsky2@skoltech.ru

Ivan Oseledets
AIRI
Moscow, Russia
oseledets@airi.net

ABSTRACT

A common assumption is that MoE routers primarily leverage semantic features for expert selection. However, our study challenges this notion by demonstrating that positional token information also plays a crucial role in routing decisions. Through extensive empirical analysis, we provide evidence supporting this hypothesis, develop a phenomenological explanation of the observed behavior, and discuss practical implications for MoE-based architectures.

1 INTRODUCTION

The integration of the Mixture of Experts (MoE) approach, originally proposed by (Jacobs et al., 1991) and (Jordan and Jacobs, 1994), into TRANSFORMER-based models has been a key driver of recent advancements in machine learning, particularly in natural language processing (NLP) (Dai et al., 2024; Muennighoff et al., 2024; Qwen et al., 2024; 2025). This innovation enables models to scale efficiently, achieving higher overall parameter counts and improved performance on downstream tasks while maintaining manageable computational requirements for training.

Mixture of Experts (MoE) approach forms a common building block which includes a set of "experts", typically neural networks of the same architecture but different weights, and a "router", a linear multi-class classifier that selects experts. The construction usually substitutes feed-forward networks in TRANSFORMERS-blocks. Only a limited subset of experts is used for processing a single input (or token), which is an appealing feature of MoE in training and inference.

General consensus is that experts are semantically specialized, and combined with dynamical gating, become an appealing approach for building large language models (LLMs). However, preliminary experimental results show significant role of token position, challenging this assumption. In this work, we address this misconception and provide multiple experiments that support our hypothesis.

Hypothesis 1.1 *(Main observation)* Mixture-of-Experts (MoE) router in TRANSFORMERS exploits positional token information in addition to semantic one.

We present experimental evidence supporting our hypothesis in Section 2. Further, we offer a phenomenological explanation and discuss empirical findings in more formal terms in Section 3. Finally, we explore the practical implications of both the empirical observations and the phenomenological model in Section 4.

2 EMPIRICAL STUDY

We begin with a preliminary experiment in Section 2.1 in order to gain a general understanding of expert activations. Next, we estimate the correlations between activations of different experts Section 2.2. Additionally, we train a classifier on embedded token sequences to predict token posi-
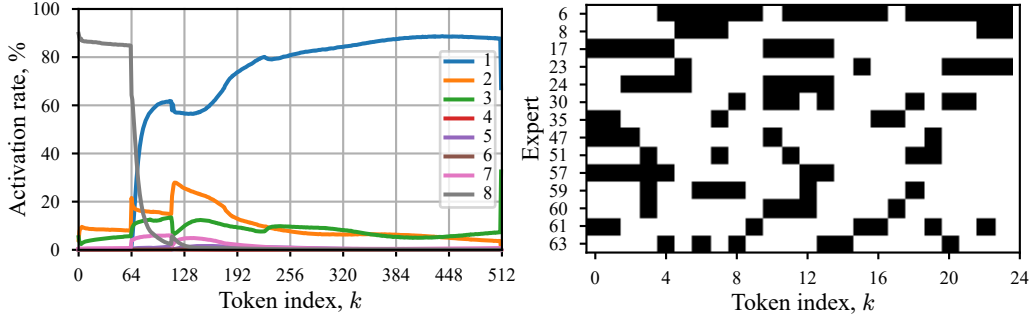
Figure 1: Average expert activation rates in the sixth MoE-block measured on SLIMPAJAMA dataset for SWITCH8 (left) and sample expert activation trajectory for OLMoE (right). The activations trajectory is sampled from the bottom layer for the first 24 tokens with first 14 most active experts.

tions in Section 2.3. This experiment evaluates the model's ability to recover token positions, which is a necessary condition for the emergence of spatial structures. See Section C for futher details.

2.1 EXPERT ACTIVATION RATE

We observe two qualitatively distinct behaviours. The first corresponds to SWITCH model (see Figure 1), where certain experts exhibit significantly higher activation rates, forming long sequences of repeatedly active experts with stable rate values. In contrast, the OLMoE model displays a different pattern (see Figure 3), with expert activation rates r_{ijk} fluctuating around an average value $\bar{r}_{ij}$. We hypothesize that the difference arises from the top-1 selection. Alternatively, it could be due to a bug in the implementation provided in TRANSFORMERS (Wolf et al., 2020).

2.2 EXPERT CORRELATION LENGTH

Our experiments show that the correlation lengths ξ_{model} of the models are larger than those of random expert activations, with a mild growth of ξ with depth of MoE-layer l, peaking in the middle layers (see Figure 2). This suggests consecutive activations of the same experts, forming patterns influenced by high-level features. Scaling of ξ_{model} with N_{block} indicates non-uniform expert activation and decaying long-range correlations. Since MoE-router is independent of token position, we suggest that experts interact with each other through embeddings vectors, which carry positional information from RoPE embeddings within the attention mechanism.

2.3 TOKEN POSITION PREDICTION

If MoE-router is capable of capturing positional token information, then a classifier of comparable complexity should be able to capture it as well. While predicting exact token position k is hard merely because of the large number of classes, the classifier should be able to predict simpler synthetic targets, such as parity $2 \mid k$ (even or odd) or the index of the subsequence of n tokens, $\lfloor \frac{k}{n} \rfloor$.

According to Table 1, MoE-router is potentially capable of extracting positional information from embeddings. While it cannot accurately predict the exact token position, it can more reliably predict the blocks index $\lfloor \frac{k}{n} \rfloor$. Interestingly, the most notable difference in predicting is between parity and block index of size 2. The classifier fails to predict parity but can successfully determine whether a token belongs to the first or second half of a sequence. This points out to the importance of low-frequency components of RoPE, while highlighting the limited utility of its high-frequency parts.

3 PHENOMENOLOGICAL MODEL

Based on the empirical observations described in Section 2, we propose a phenomenological model in Section 3.1 that characterizes the MoE-model from the perspective of statistical mechanics (SM).
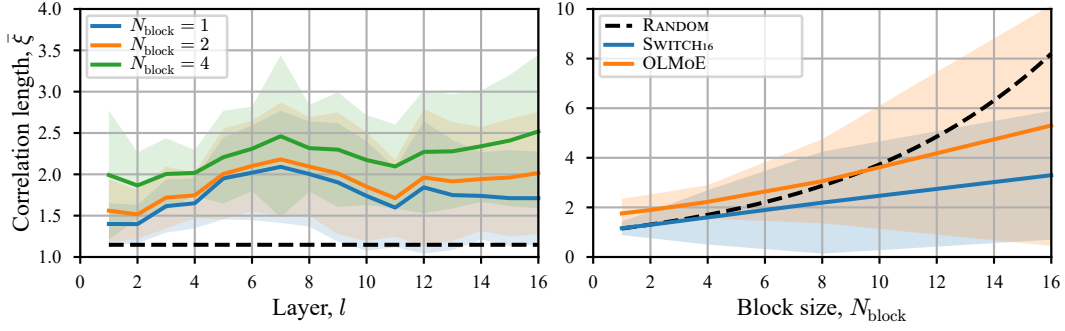
Figure 2: Averaged correlation length $\bar{\xi}$ scaling. Correlation lengths are averaged over experts in OLMoE layers (left) or in the entire model (right).

Subsequently, we address the load balancing problem and suggest a new auxiliary loss within the proposed model in Section 3.2.

3.1 STATISTICAL MODEL

Consider a sequence of tokens t_i and corresponding input embeddings x_i with $0 \leq i < L-1$. Each token t_i can be attributed to one of the states s_i. Gibbs gives a probability distribution on states

$$p(s_i|\Theta, t_1, t_2, ..., t_{L-1}) \propto e^{-\beta E(s_i; \Theta, t_1, t_2, ..., t_{L-1})}, \tag{1}$$

where set of all tokens t_i parametrizes energy function E (we omit tokens in the parameter specification from now on) and $\beta = T^{-1}$ is inversed temperature. Natural ordering of tokens and corresponding experts forms a one dimensional lattice (or chain). This chain of experts (spins in SM) defines a one dimensional Ising model of a mixture of experts governed by Hamiltonian $H(s_i)$.

The form of Hamiltonian E essentially defines all properties of the model. However, it is quite difficult to provide closed analytical expression for E but it is sufficient enough to reason about main characteristics and spin interactions in particular. For example, an attention mechanics ensures non-linear expert-expert interaction but its range is a more subtle topic.

We assume that a typical MoE model like OLMoE (Muennighoff et al., 2024) demonstrates a short-range expert-expert interactions. From general consideration, one may expect that a TRANSFORMER for language modeling tends to operate on a near context of tens or hundreds of tokens. For example, grammar and syntax require agreement and concordance among words in an utterance.

The more specific argument is based on decaying of attention scores. It is a complex and different subject which goes beyond this work and which requires an additional study. However, modern TRANSFORMER-based architectures and models used for empirical study in Section 2 operate internally with a variant of relative positional encodings (Shaw et al., 2018) which admit scores decaying with relative distance between tokens. Specifically, RoPE positional encodings (Su et al., 2022), a building block of OLMoE, do indeed decays (see Section 3.4.3 in (Su et al., 2022)).[1] This constitutes the rationale for using one dimensional model defined above as a phenomenological model of MoE-blocks in TRANSFORMERS.

3.2 LOAD BALANCING PROBLEM

The load balancing problem is a problem of assigning equisized batches of tokens to each expert (Fedus et al., 2022a). Without extra efforts, a few most active expert creates a positive feedback loop during training that makes them the only active experts in training and inference time. This well-known fact can be reframed with our spin-glass model.

High activation rates of some experts corresponds high density of these experts (spins) in our spin-glass. Experts can appear in different spatial structures forming topologically ordered phases.

---

[1]More precisely, Su et al. (2022) give only an upper bound on the multiplication factor. Long-range decay of RoPE is also challenged by Barbero et al. (2024).

Table 1: Quality metrics of token position k prediction task against different (syntetic) targets for the first MoE-block in OLMoE. Token position parity is $2 \mid k$. Target for classification on consecutive blocks of size n is denoted as $\lfloor k/n \rfloor$. Shadowed value in parentheses denotes standard deviation.

TARGET	CLASSES	ACC@1	ACC@2	ACC@8	AP	PR	RECALL	F1
$2 \mid k$	2	49.9(5)	—	—	50.1(7)	49.9(5)	50.4(4)	50.1(4)
$\lfloor k/128 \rfloor$	2	91.7(3)	—	—	96.4(5)	90.0(3)	93.9(3)	91.9(3)
$\lfloor k/64 \rfloor$	4	75.9(3)	95.9(2)	—	82.8(3)	76.0(3)	75.9(3)	75.9(3)
$\lfloor k/16 \rfloor$	16	41.8(1)	64.7(2)	96.2(2)	41.1(4)	41.3(1)	41.8(1)	41.3(1)
$\lfloor k/4 \rfloor$	32	13.8(2)	24.4(5)	59.5(6)	11.8(1)	13.5(2)	13.8(2)	13.6(2)
$\lfloor k/2 \rfloor$	128	6.8(3)	12.4(5)	35.2(1)	5.8(1)	6.73(2)	6.8(3)	6.7(3)
k	256	3.6(1)	6.0(1)	17.4(3)	3.0(1)	3.6(2)	3.6(1)	3.5(2)

However, Landau's argument for absence of ordering in one dimension (Landau and Lifshitz, 2013) breaks any ordering at all.

Remark 3.1. In contrast to one dimensional Ising model, the general case of $n > 1$ allows existence of ordering. For example, two dimensional lattices of experts and tokens can emerge from models with augmented context like RETRO (Borgeaud et al., 2022). This particularly means that training of such kind of models could potentially require additional effort in respect to conventional MoE-models. On the other hand, the absence of ordering in one dimension could imply reduced model expressivity since some expert configurations are topologically prohibited.

The only remaining condition that must be met is equilibrium state, i.e. entropy $H(p)$ is maximal.

$$\max_{\Theta} \mathcal{L}_{\mathrm{MEM}}(p; \Theta), \quad \mathcal{L}_{\mathrm{MEM}}(p; \Theta) = T \sum_{i=1}^{L} H(p(s_i)). \tag{2}$$

In machine learning literature, this method is known as maximum-entropy principle (Jaynes, 1957a; 1957b). Obviously, our MEM-loss should be taken with negative sign in order to be used as an auxiliary loss term for end-to-end training of entire model. Optimization problem (2) can be rewritten as a minimization problem of KL-divergence between the proposed distribution and the equilibrium:

$$\min_{\Theta} \; T \sum_{i=1}^{L} D_{\mathrm{KL}}(p(s_i; \Theta) \parallel q). \tag{3}$$

Upper bound of $H(p)$ for discrete states s_i corresponds to uniform distribution on s_i with probability q_i. Since there are $\binom{n}{k}$ different states of k active experts out of n experts, $q_i = 1/\binom{n}{k}$. For example, $q = 1/\binom{64}{8} \approx 2.26 \cdot 10^{-10}$ in case of OLMoE.

4 PRACTICAL IMPLICATIONS

Observations made in Section 2 lead to several practical implications which we discuss here.

Observation 4.2 *(Spatial structure)* Figure 2 and Table 1 suggest that a spatial correlation among experts exists and experts tend to form spatial structures.

Observation 4.2 motivates study of a static routing as alternative to the dynamic gating mechanism in MoE. Static routing means an expert activation in dependence on its position. It has multiple potential advantages in comparison to dynamic gating. First, static routing is less sensitive to data shuffling and uneven communications among model shards with `scatter` and `gather` primitives. Second, training complications like load balancing heuristics and loss terms can be neglected.

Observation 4.3 Statistical mechanics suggests a valid phenomenological model of MoE.

Observation 4.3 supports the applicability of the entropy maximization principle (2). This results in the formulation of the MEM-loss (3) as an alternative to aux-loss (Fedus et al., 2022a) or Z-loss

(Zoph et al., 2022). MEM-loss offers a clear interpretation and is theoretically grounded. However, its practical validation in training is left for future work.

5 COMMENTS AND DISCUSSION

In this work, we studied internal MoE dynamics empirically. Specifically, we found and experimentally demonstrated spatial correlations in expert activations. Token position prediction experiment highlights importance of positional information for entire TRANSFORMER architecture. The phenomenological model provides a perspective of statical mechanics and motivates MEM-loss, a theoretically grounded alternative to load balancing loss. Training MoE-model from scratch with MEM-loss and experimenting with bigger models and models of different architectures are left for future work.

REFERENCES

Barbero, F., Vitvitskyi, A., Perivolaropoulos, C., Pascanu, R., and Veličković, P. *Round and Round We Go! What makes Rotary Positional Encodings useful?*. arXiv, 2024, October 8. https://doi.org/10.48550/arXiv.2410.06205

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. B. V. D., Lespiau, J.-B., Damoc, B., Clark, A., Casas, D. D. L., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., … Sifre, L. Improving Language Models by Retrieving from Trillions of Tokens. *Proceedings of the 39th International Conference on Machine Learning*, 2206–2240, 2022. https://proceedings.mlr.press/v162/borgeaud22a.html

Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y. K., Huang, P., Luo, F., Ruan, C., Sui, Z., and Liang, W. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *CoRR*, 2024. https://openreview.net/forum?id=3FQRs7iDVa

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M. P., Zhou, Z., Wang, T., Wang, E., Webster, K., Pellat, M., Robinson, K., … Cui, C. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. *Proceedings of the 39th International Conference on Machine Learning*, 5547–5569, 2022. https://proceedings.mlr.press/v162/du22c.html

Fedus, W., Dean, J., and Zoph, B. *A Review of Sparse Expert Models in Deep Learning*. arXiv, 2022b, September 4. https://doi.org/10.48550/arXiv.2209.01667

Fedus, W., Zoph, B., and Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, *23*(120), 1–39, 2022a. http://jmlr.org/papers/v23/21-0998.html

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive Mixtures of Local Experts. *Neural Computation*, *3*(1), 79–87, 1991. https://doi.org/https://doi.org/10.1162/neco.1991.3.1.79

Jaynes, E. T. Information Theory and Statistical Mechanics. *Physical Review*, *106*, 620–630, 1957a. https://doi.org/10.1103/PhysRev.106.620

Jaynes, E. T. Information Theory and Statistical Mechanics. II. *Physical Review*, *108*, 171–190, 1957b. https://doi.org/10.1103/PhysRev.108.171

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. d. l., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., … Sayed, W. E. *Mixtral of Experts*. arXiv, 2024, January 8. https://doi.org/10.48550/arXiv.2401.04088

Jiang, W. The VC Dimension for Mixtures of Binary Classifiers. *Neural Comput.*, *12*(6), 1293–1301, 2000. https://doi.org/10.1162/089976600300015367

Jordan, M. I., and Jacobs, R. A. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, *6*(2), 181–214, 1994. https://doi.org/10.1162/neco.1994.6.2.181

Kang, K., and Oh, J.-H. Statistical Mechanics of the Mixture of Experts. *Advances in Neural Information Processing Systems*, *9*, 1996. https://proceedings.neurips.cc/paper/1996/hash/a7d8 ae4569120b5bec12e7b6e9648b86-Abstract.html

Landau, L., and Lifshitz, E. *Theoretical Physics: Statistical Physics* (Vol. 5, Issue 1). Butterworth-Heinemann, 2013. https://books.google.ru/books?id=VzgJN-XPTRsC

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. *GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding*. International Conference on Learning Representations, 2020, October 2. https://openreview.net/forum?id=qrwe7XHTmYb

Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., Shi, W., Walsh, P., Tafjord, O., Lambert, N., Gu, Y., Arora, S., Bhagia, A., Schwenk, D., Wadden, D., Wettig, A., Hui, B., Dettmers, T., Kiela, D., … Hajishirzi, H. *OLMoE: Open Mixture-of-Experts Language Models*. arXiv, 2024, September 3. https://doi.org/10.48550/arXiv.2409.02060

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830, 2011.

Qwen, Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., … Fan, Z. *Qwen2 Technical Report*. arXiv, 2024, September 10. https://doi.org/10.48550/arXiv.2407.10671

Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., … Qiu, Z. *Qwen2.5 Technical Report*. arXiv, 2025, January 2. https://doi.org/10.48550/arXiv.2412.15115

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1–67, 2020. http://jmlr.org/papers/v21/20-074.html

Rajbhandari, S., Li, C., Yao, Z., Zhang, M., Aminabadi, R. Y., Awan, A. A., Rasley, J., and He, Y. DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale. *Proceedings of the 39th International Conference on Machine Learning*, 18332–18346, 2022. https://proceedings.mlr.press/v162/rajbhandari22a.html

Shaw, P., Uszkoreit, J., and Vaswani, A. Self-Attention with Relative Position Representations. In M. Walker, H. Ji, & A. Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468. Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/N18-2074

Soboleva, D., Al-Khateeb, F., Myers, R., Steeves, J. R., Hestness, J., and Dey, N. *SlimPajama: A 627B token cleaned and deduplicated version of RedPajama*, 2023. https://huggingface.co/datasets/cerebras/SlimPajama-627B

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., and Liu, Y. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. arXiv, 2022, August 8. https://doi.org/10.48550/arXiv.2104.09864

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. *Transformers: State-of-the-Art Natural Language Processing*. 38–45, 2020. https://www.aclweb.org/anthology/2020.emnlp-demos.6

Xue, F., Zheng, Z., Fu, Y., Ni, J., Zheng, Z., Zhou, W., and You, Y. OpenMoE: An Early Effort on Open Mixture-of-Experts Language Models. *Proceedings of the 41st International Conference on Machine Learning*, 55625–55655, 2024. https://proceedings.mlr.press/v235/xue24c.html

Yuksel, S. E., Wilson, J. N., and Gader, P. D. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), 1177–1193, 2012. https://doi.org/10.1109/TNNLS.2012.2200299

Zhu, T., Qu, X., Dong, D., Ruan, J., Tong, J., He, C., and Cheng, Y. LLaMA-MoE: Building Mixture-of-Experts from LLaMA with Continual Pre-Training. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15913–15923. Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.890

Zoph, B., Bello, I., Kumar, S., Du, N., Huang, Y., Dean, J., Shazeer, N., and Fedus, W. *ST-MoE: Designing Stable and Transferable Sparse Expert Models*. arXiv, 2022, April 29. https://doi.org/10.48550/arXiv.2202.08906

A   RELATED WORKS

Comprehensive surveys of Mixture of Experts (MoE) can be found in (Yuksel et al., 2012) and (Fedus et al., 2022b). The first work (Yuksel et al., 2012) focuses on early works in the first twenty years of MoE development and the second one (Fedus et al., 2022b) cover the next ten years. Initially, MoE approach has been described in (Jacobs et al., 1991) then it has been generalized to broaded class of hierarchical models in (Jordan and Jacobs, 1994).

Combinatorical properties. In work (Jiang, 2000) authors studed combinatorial properties of MoE models. Specifically, they provided bounds for mixture of Bernoulli classifier and mixture of logistic regression classifier.

Statistical properties. In work (Kang and Oh, 1996) authors considered MoE models from statistical mechanics.

B   AVAILABLE MoE-TRANSFORMERS

In all experiments, we use only pretrained models published on HUGGINGFACE HUB. We make a long list of recent TRANSFORMERS models with MoE-adapter presented in literature (Dai et al., 2024; Du et al., 2022; Jiang et al., 2024; Lepikhin et al., 2020; Muennighoff et al., 2024; Rajbhandari et al., 2022; Xue et al., 2024; Zhu et al., 2024; Zoph et al., 2022) (see Table 2). Despite the fact there are plenty of models available online, we are limited by two factors. Firstly, only a few models have a necessary instrumentation required to get MoE router output logits and selected experts. The second limitation stems from hardware available to us (i.e. 2x Nvidia V100). It's getting even worse with the fact that the majority of models use bfloat16 for arithmetics which has no native support on GPUs of Volta family (fortunately, emilation via fp32 rescues the day).

We use PYTORCH for our experiments. For model inference, autodiff is disable with `torch.inference_mode()` context but models stay in training mode (i.e `.training` attribute evaluates to true). In this way, we save memory for larger batch and sample expert activations as if we indeed train a model.

C   MISCELLANEOUS EMPIRICAL STUDY

In order to ensure versatility among model architectures as well as MoE layer architectures, we consider MoE-models of distinct two architectures: SWITCH (Fedus et al., 2022a) and OLMoE (Muennighoff et al., 2024). SWITCH is an encoder-decoder TRANSFORMER based on T5 (Raffel et al., 2020) with a single active expert and expert capacity $C_{\text{expert}} = 64$. We use its two variants with total number of experts $N_{\text{expert}} = 8$ and $N_{\text{expert}} = 16$ (SWITCH8 and SWITCH16). Despite the encoder-decoder nature of SWITCH, we pass an empty premise to encoder. OLMoE is a decoder-only transformer with 8 active experts out of total $N_{\text{expert}} = 64$ experts. Pre-trained model checkpoints published on the HUGGINGFACE HUB were used in all experiments (see Section B). In addition, we use a test split of a diverse SLIMPAJAMA corpus (Soboleva et al., 2023) to guarantee the representativeness samples of expert activations.
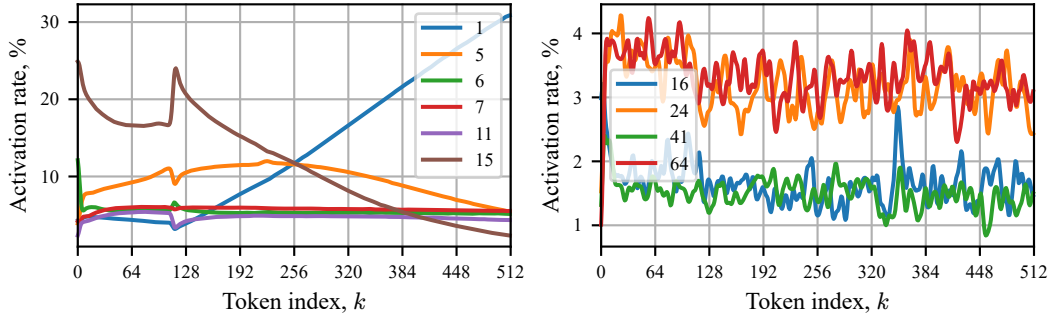
C.1   EXPERT ACTIVATION RATE

Figure 3: Expert activation rates for the first 512 tokens and some experts of SWITCH₁₆ (left) and OLMoE (right) models. Expert choice is purely random with a minor exception. Experts 1 and 15 of SWITCH₁₆ are outliers and demonstrates atypical behavior. Expert activation rates of OLMoE are smoothed for better representation with Gaussian filter of $\sigma = 2$.

In order to measure in what degree experts are interleaved or overlapped each other, we collect expert activation frequencies for specific token position and TRANSFORMER-block on a sample sequences. Sequence are sampled from test split of SLIMPAJAMA corpus with about 500k documents. Then aggregated frequencies over all samples c_{ijk} are used to estimate expert activation rates as $r_{ijk} = c_{ijk} / \sum_k c_{ijk}$ (see Figure 1 and 3).

C.2 EXPERT CORRELATION LENGTH

Table 2: Incomplete list of availabel MoE models that can potentially be used in for experimentation.

MODEL	REFERENCE	PARAMETERS	HUGGINGFACE (🤗)
GShard	(Lepikhin et al., 2020)	37B	—
DeepSpeed-MoE	(Rajbhandari et al., 2022)	350M/13B	—
		PR-350M/4B	—
		PR-1.3B/31B	—
GLAM	(Du et al., 2022)	0.1B/1.9B	—
		1.7B/27B	—
ST-MoE	(Zoph et al., 2022)	0.8B/4.1B	—
Switch	(Fedus et al., 2022a)	250M	google/switch-base-8
		1B	google/switch-base-16
		2B	google/switch-base-32
		4B	google/switch-base-64
DeepSeek-MoE	(Dai et al., 2024)	0.24B/1.89B	—
		2.8B/16.4B	deepseek-ai/deepseek-moe-16b-base
LLaMA-MoE	(Zhu et al., 2024)	3.0B/6.7B	llama-moe/LLaMA-MoE-v1-3_0B-2_16
		3.5B/6.7B	llama-moe/LLaMA-MoE-v1-3_5B-4_16
		3.5B/6.7B	llama-moe/LLaMA-MoE-v1-3_5B-2_8
Mixtral	(Jiang et al., 2024)	13B/47B	mistralai/Mixtral-8x7B-v0.1
		39B/141B	mistralai/Mixtral-8x22B-v0.1
OLMoE	(Muennighoff et al., 2024)	1B/7B	allenai/OLMoE-1B-7B-0924
OpenMoE	(Xue et al., 2024)	8B	OrionZheng/openmoe-8b

Table 3: Fitted parameters of law $\xi/\xi_0 = \exp(\alpha n)$ to experimental data.

MODEL	N_{expert}	α	ξ_0
RANDOM	1/8	0.131269	0.005969
	1/16	0.131088	0.006606
OLMoE	8/64	0.073877	0.499299
SWITCH16	1/16	0.068579	0.146133

We use two definitions of correlation length. The first one ξ_{dw} is defined as the ratio of the sequence length L to the average number of domain walls $\bar{N}$. The second one ξ_{ds} is estimated with direct computation of domain sizes L_{e} with subsequent averaging.

$$\xi_{\text{dw}} = \frac{L}{\bar{N}_{\text{e}}}, \quad \xi_{\text{ds}} = \bar{L}_{\text{e}}. \tag{4}$$

In practice, ξ_{dw} and ξ_{ds} are correlated. However, ξ_{ds} is more precise definition but more costly to estimate than ξ_{dw}. Henceforth, ξ is used to denote ξ_{ds} without label.

Domain size is estimated over block variables, i.e. block of N_{block} sequential expert activation indicators. If an expert k is activated in merely one indicator of a block, than the entire block indicates activation of expert k (i.e. disjunctive union of experts).

From analysis of experimental data, we find that ξ scales well with n according exponential law, i.e.

$$\frac{\xi}{\xi_0} = e^{\alpha n}. \tag{5}$$

We fit exponents α for models of interest and present them in Table 3.

C.3 TOKEN POSITION PREDICTION

We dissect OLMoE and keep only the first layer up to MoE-router. We sample embeddings short sequences of length $L = 256$ right before MoE-router but after layer normalization (normalization is critical for training a linear models). Finally, we train a multinomial logistic with SCIKIT-LEARN (Pedregosa et al., 2011) with stratified 3-fold cross validataion and grid search over L_2 reguralizer.