Scaling Fine-Grained MoE Beyond 50B Parameters: Empirical Evaluation and Practical Insights

Jakub Krajewski [1 2] **Marcin Chochowski** [3] **Daniel Korzekwa** [3]

Abstract

Mixture of Experts (MoE) architectures have emerged as pivotal for scaling Large Language Models (LLMs) efficiently. While fine-grained MoE approaches - utilizing more numerous, smaller experts - have shown promise, quantification of their efficiency gains at large scales remains crucial. This work proposes a set of training recipes and provides a comprehensive empirical evaluation of fine-grained MoE, directly comparing its scaling properties against standard MoE configurations for models with up to 56B total (17B active) parameters. We investigate convergence speed, model performance on downstream benchmarks, and practical training considerations across various setups. Overall, at the largest scale we observe that fine-grained MoE achieves better validation loss and higher accuracy across a set of downstream benchmarks. This study offers empirical grounding and practical insights for leveraging fine-grained MoE in the development of future large-scale models.

1. Introduction

Extraordinary capabilities shown in recent years by Large Language Models (LLMs) (OpenAI et al., 2024; Team et al., 2024) come with steep resource requirements, prompting the search for more cost-effective training methods (Faiz et al., 2024; Touvron et al., 2023a). Among such approaches, Mixture of Experts (MoE) (Shazeer et al., 2017) has emerged as a particularly successful technique, activating only a fraction of the total parameters for each input token. Building upon the standard MoE architecture, fine-grained Mixture of Experts (Dai et al., 2024; Krajewski et al., 2024) leverages a larger number of smaller experts and routes tokens to multiple experts at once, maintaining computational efficiency while often yielding faster convergence.

[1]IDEAS NCBR [2]University of Warsaw [3]NVIDIA. Correspondence to: Jakub Krajewski <gim.jakubk@gmail.com>.

This paper proposes a set of training recipes to effectively scale fine-grained MoE up to 56B total model parameters. We conduct a large-scale empirical evaluation comparing this approach against two standard MoE variants: Top-1 (Switch (Fedus et al., 2022)) and Top-2 (GLaM (Du et al., 2022), Mixtral (Jiang et al., 2024)). All of the experiments use the same controlled setup, including dataset and evaluation protocol, allowing for a fair comparison. We consider multiple model sizes and training durations, allowing to compare the results on different scaling axes. We measure both the pretraining loss convergence speed and downstream accuracy on a set of popular benchmarks.

Our key contributions are summarized as follows:

- We propose recipes and compare the convergence of the models across different model and dataset sizes, scaling up to 56B total (17B active) parameters. On the largest scale, we show gains from increasing granularity in both baseline configurations (Fig. 1 (d)-(e)).
- Beyond perplexity-based comparison, we evaluate the models on a set of benchmarks. We demonstrate that the improvements in pretraining loss transfer to downstream accuracy (Table 1 & Table 3).
- We analyze and discuss hyperparameter choices and training specifics of fine-grained MoE models (Sec. 4). Specifically, we highlight the importance of the order of softmax and Top-k normalization in the router, evaluate the expert load imbalance, and analyze the evolution of router logits throughout training.

2. Background

Below we present background on fine-grained MoE, while a more detailed Related Work section can be found in App. F.

(Dai et al., 2024; Krajewski et al., 2024) propose to relax the assumption that each MoE expert is the same size as the standard Feed-Forward layer. Instead, we can use more, smaller experts, increasing the flexibility of mapping tokens to experts. For the Transformer with a hidden Feed-Forward size d_{ff}, in the standard MoE layer we will choose k experts for each token, with the expert hidden size $d_{\text{expert}} = d_{\text{ff}}$, and

total number of experts N_E. Then, for fine-grained MoE with granularity G, we increase the total number of experts to $G \cdot N_E$, decrease the expert hidden size to $d_{\text{expert}} = d_{\text{ff}}/G$, and choose $G \cdot k$ experts for each token. This way, the non-router FLOPs and parameters remain unchanged.

Using granularity $G > 1$ is the optimal choice from the scaling laws perspective (Krajewski et al., 2024). Furthermore, architectures using fine-grained MoE have shown superior performance (DeepSeek-AI et al., 2024a;b). However, they often also use other architecture advancements, potentially contributing to the overall efficiency improvement. To better understand the scaling advantage of fine-grained MoE, in this report we present the details of our ablations and pretraining experiments on up to 56B of total parameters. Additionally, we show downstream evaluations, a case which was not considered in (Krajewski et al., 2024), who only focused on modeling perplexity.

3. Scaling and Evaluation of Fine-Grained MoE

3.1. Experimental setup

Below we describe the considered MoE architectures and outline the training framework and hardware used in the experiments. Full details on the training hyperparameters, dataset and procedure can be found in App. A.

We compare four MoE variants: two common baselines and their fine-grained counterparts. In each pair, the standard and fine-grained models are matched in parameters and compute (excluding router FLOPs, negligible for our models):

- A Switch-like MoE (Fedus et al., 2022) with 8 experts and Top-1 routing (denoted as 1×FLOPs-G1), and a fine-grained version with 64 experts and Top-8 routing (1×FLOPs-G8).
- A Mixtral-like MoE (Jiang et al., 2024) with 8 experts and Top-2 routing (2×FLOPs-G1), and its fine-grained counterpart with 64 experts and Top-16 routing (2×FLOPs-G8).

These fine-grained models correspond to granularity 8 as defined in Krajewski et al. (2024). We find setting $G = 8$ sufficient to compare standard and fine-grained MoEs. Higher G would likely further amplify the observed gains, but could also be more challenging to implement efficiently (Tan et al., 2024). We use the notation 1xFLOPs/2xFLOPs referring to the FLOPS *in the MoE layer*: the latter variants activate twice the number of experts per token, hence proportionally increasing the computational cost. The models presented in this report have been trained on a cluster featuring NVIDIA H100 GPUs. All experiments are run using the open-source Megatron-LM framework (Shoeybi et al., 2020).

3.2. Comparing 11B Models

We begin our comparison by examining the effects of granularity and MoE design on models with approximately 11B total (3B-4B active) parameters, pretrained on 50B tokens. Detailed model configurations are listed in Table 5 (App. A). Fig. 1 (a)-(b) depicts the loss curves of the models, while Table 1 presents the downstream benchmark results for these four configurations (full list of all scores can be found in App. B). Below we analyze the impact of granularity within each FLOPs category (1x and 2x).

Comparing the Switch-style variants, we observe that adding granularity (1xFLOPs-G8) significantly improves accuracy over the standard configuration (1xFLOPs-G1). This result proves the advantage of using fine-grained experts in this setup.

As expected, the standard Mixtral-like model (2xFLOPS-G1), activating two experts per token, outperforms the standard Switch-like model (1xFLOPS-G1), which activates only one. This reflects the benefit of increasing the number of parameters activated per each token.

Interestingly, when comparing the 2xFLOPs variants, adding granularity (2xFLOPS-G8) does not yield a similarly significant improvement over the standard configuration (2xFLOPS-G1). The validation losses and benchmark scores of both models are very close. In Sec. 3.3 & Sec. 3.4 we show that this observation changes with sufficiently long training and larger models. We offer a possible explanation of this phenomenon in Sec. 4.

Table 1. Downstream evaluation scores for the 11B models. Within each pair (Switch-like, Mixtral-like), we mark the model with a better score for each metric. See App. B for results on all considered benchmarks.

MODEL	ARC-E	ARC-C	PIQA	SIQA
1xFLOPs-G1	58.1	29.7	71.9	41.6
1xFLOPs-G8	**60.1**	**32.7**	**73.0**	**42.6**
2xFLOPs-G1	60.6	31.4	**75.0**	42.6
2xFLOPs-G8	**62.3**	**32.8**	74.2	**44.3**

3.3. The Effect of Training Length

Modern LLMs are pretrained on very large datasets and understanding how the training length affects different MoE configurations can provide an important context to our analysis. To examine scaling the dataset size, we perform additional experiments training the 11B models on two token horizons - 25B and 100B tokens - alongside the original setup. Fig. 1(c) depicts the validation losses for these three training lengths. Notably, we observe more pronounced performance gains of fine-grained MoE as the dataset size grows.

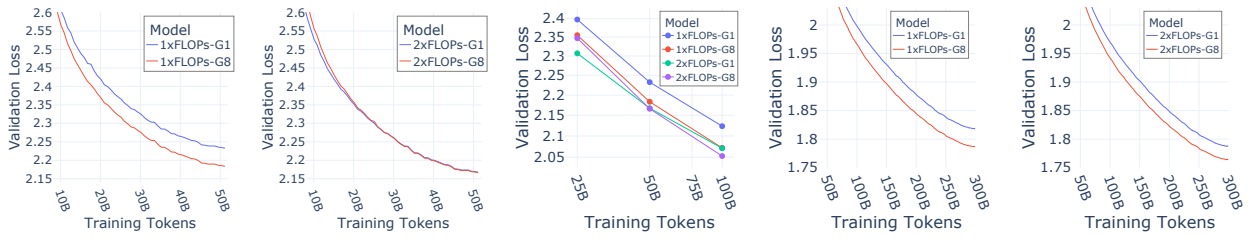


Figure 1. **(a)** - **(b)**: Validation loss curves of 11B models. Granularity improves performance of the Switch model, but doesn't bring advantage for the Mixtral variant. This observation changes with longer training. **(c)**: Comparison of the final loss for scaling the training length of the 11B models. Fine-grained variants perform relatively the best on the longest token horizon. **(d)** - **(e)**: Validation loss curves of 56B models. Granularity brings improvement in both Switch-like and Mixtral-like setup.

At the shorter horizon of 25B tokens, the two fine-grained variants show validation losses similar to each other and fall between the two standard baselines. Surprisingly, the 2xFLOPs-G8 variant offers only a very modest improvement over 1xFLOPs-G8, even though it activates twice as many experts per token.

This picture changes when comparing models on the longest token horizon. In that setup, 1xFLOPs-G8 matches 2xFLOPs-G1. Notably, this means that the 1xFLOPs-G8 model achieves similar performance to 2xFLOPs-G1, despite activating just about half the MoE parameters. This results in cheaper training, but also a more lightweight inference. The 2xFLOPs-G8 model outperforms its non-granular counterpart, 2xFLOPs-G1.

Table 2. Training step savings measured in % of training steps needed to reach the final loss of the standard Switch-MoE (1xFLOPs-G1). The savings increase for fine-grained MoE and remain constant for the standard Mixtral architecture.

Model	Training Tokens		
	25B	50B	100B
1xFLOPs-G8	21.6%	27.9%	33.6%
2xFLOPs-G1	34.5%	32.8%	32.8%
2xFLOPs-G8	24.6%	32.8%	39.4%

Directly comparing absolute pretraining loss values can be difficult. To provide a more intuitive measure of efficiency gains, in Table 2 we report *training step savings*. This metric quantifies how many fewer training steps a specific variant needs to reach the same final validation loss as the baseline Switch model. We report savings as a percentage of the total training steps. Examining the results, we can see that the standard Mixtral model shows a roughly constant advantage among the three dataset sizes. In contrast, the savings from using fine-grained MoE are growing with the training duration. This suggests that the granular models become increasingly efficient with more training data.

3.4. Scaling Beyond 50B Parameters

We further scale our analysis by comparing similar set of models this time with 56B total (11B-17B active) parameters, trained on 300B tokens. The key architecture hyperparameters are presented in Table 6 (App. A).

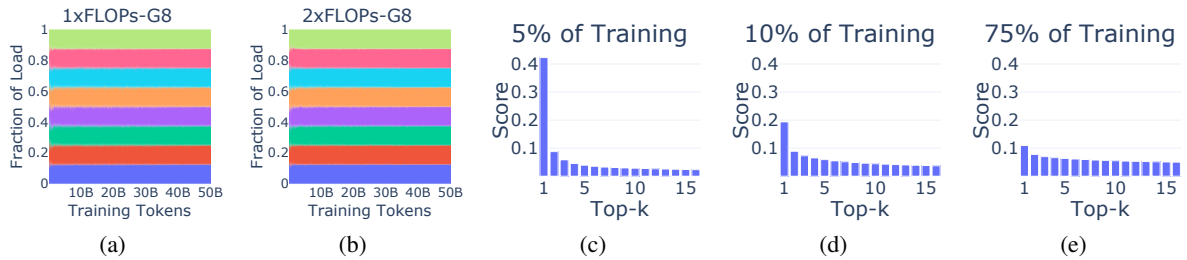
Loss curves of the models are shown in Fig. 1 (d)-(e). We observe the improvement from using fine-grained MoE in both Switch and Mixtral variants. Consequently, the relative performance ranking of the considered architectures closely resembles the order observed for the 11B models on the longest token horizon (Sec. 3.3). The 1xFLOPs-G8 variant outperforms the standard Switch MoE and matches the more computationally expensive 2xFLOPs-G1. Adding granularity on top of the Mixtral variant also results in a clear improvement.

Table 3 shows how these observations translate to downstream evaluations (full list of scores on all considered benchmarks can be found in App. B). We see the same relationships in performance as observed for validation loss.

Scaling to 56B parameters confirms the benefits of fine-grained MoE for both Switch and Mixtral variants. Notably, the 1xFLOPs-G8 model matches the performance of the standard 2xFLOPs-G1 while activating fewer parameters in each forward pass. The 2xFLOPs-G8 model achieves the best overall results, further supporting the effectiveness of fine-grained MoE at larger scales.

Table 3. Downstream evaluation scores for the 56B models. Within each pair (Switch-style, Mixtral-style), we mark the model with a better score for each metric. See App. B for all benchmarks.

Model	MMLU	ARC-E	ARC-C	PIQA
1xFLOPs-G1	47.5	73.2	44.3	80.5
1xFLOPs-G8	**50.1**	**76.1**	**46.4**	**81.0**
2xFLOPs-G1	50.5	74.3	45.6	79.9
2xFLOPs-G8	**52.1**	**77.6**	**47.7**	**81.2**



(a) (b) (c) (d) (e)

Figure 2. **(a)** - **(b)**: Fraction of load assigned to each Expert Parallel group in the initial layer of fine-grained models. We generally observe a balanced assignment of tokens between different EP groups. This is similar in the middle and final layer of the model (see App. E). **(c)**-**(e)**: The distribution of the Top-k router logits in the initial layer of 2xFLOPs-G8 at selected points of training. In the beginning, the router assigns most of the weight to single expert, while learning to use the remaining ones later in training. We observe a similar pattern in the middle and last layers of the model and in the 1xFLOPs-G8 variant (see App. D).

4. Discussion on MoE Design and Training

Expert Load Imbalance. In practice, training MoE models often involves using Expert Parallelism (EP), distributing experts across different devices (e.g., GPUs). A crucial goal in such distributed settings is to ensure each device processes roughly the same number of tokens per forward pass, avoiding delays caused by waiting for the most heavily loaded device. The mapping of experts to devices is determined by the Expert Parallel (EP) size. For our setup, using a fine-grained MoE with 64 experts and an EP size of 8 means each device is responsible for 8 experts. To assess how well the load was balanced across devices in our fine-grained models, we tracked the fraction of the total token load processed by each of the 8 EP groups during training. These measurements are shown for the initial layer in Fig. 2 (a)-(b), and for the middle and last layers in Figs. 5 & 6 (App. E). The data generally indicates that the load balancing loss quickly leads to uniform load distribution across the EP groups. Therefore, we did not encounter significant load balancing problems with fine-grained MoE in our setup.

Evolution of the Router Logits Magnitude. As shown in Section 3.3, the performance advantage of fine-grained MoE models increases with the amount of training data. While fine-grained MoE eventually outperforms its standard counterparts, its relative gain is less pronounced on shorter training horizons. Here, we explore a potential reason for this effect. We examine the distributions of the median router logits throughout training (example shown in Fig. 2(c)-(e), more results plotted in App. D) for both 1xFLOPs-G8 and 2xFLOPs-G8 models. We can observe that early in training, the router heavily favors the top-1 expert and only gradually learns to utilize additional experts per token over time. This behavior helps to explain why the advantage of fine-grained models becomes more significant with a longer training – the router requires sufficient training data to learn how to effectively distribute load across multiple fine-grained experts.

The Order of Softmax and Top-k. The optimal ordering of softmax and Top-k normalization in the MoE router is not clear in the literature. While many works (Shazeer et al., 2017; Zoph et al., 2022; Jiang et al., 2024) put the softmax after the Top-k choice, others (Fedus et al., 2022; Xue et al., 2024; Muennighoff et al., 2024) perform the Top-k choice after softmax. Please note that to preserve gradient in the router, the latter option is necessary if choosing only 1 expert for each input; both are possible when more than 1 expert is selected. We evaluate the impact of this ordering in our experiments. As presented in Table 4, for the standard MoE model (G1), we observe similar performance between the two approaches (slightly better when applying softmax after Top-k). However, for fine-grained MoE (G8), applying softmax after the Top-k selection yields significantly better results. Consequently, throughout this paper, we use the softmax after Top-k ordering for all models where $k > 1$ (1xFLOPS-G8, 2xFLOPS-G8, 2xFLOPS-G1). For the $k = 1$ model (1xFLOPS-G1), we necessarily apply softmax before the Top-k selection to maintain gradient flow.

Table 4. Validation loss for training models with different orders of softmax and Top-k: softmax before Top-k (pre-softmax) and softmax after Top-k (post-softmax).

MODEL	PRE-SOFTMAX	POST-SOFTMAX
1×FLOPs-G8	2.219	**2.183**
2×FLOPs-G1	2.175	**2.168**
2×FLOPs-G8	2.194	**2.166**

5. Conclusions

In this report, we empirically evaluated several MoE variants, and proposed training recipes for scaling models up to 56B total (17B active) parameters. Our results show that fine-grained MoE improves model performance and computational efficiency compared to standard architecture. These findings provide empirical and practical grounding for training future large-scale fine-grained MoE models.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

We thank Ethan He for his valuable comments and suggestions throughout this project. We are grateful to Janusz Lisiecki and to Jacek Staszewski and Małgorzata Ciechomska for their instrumental support in establishing the collaboration between NVIDIA and IDEAS NCBR. We also extend our thanks to Maciej Pióro, Sebastian Jaszczur and Jan Ludziejewski for their helpful feedback on this report. Furthermore, we wish to express the gratitude to Piotr Sankowski for the creation of the supportive scientific environment.

References

Abnar, S., Shah, H., Busbridge, D., Ali, A. M. E., Susskind, J., and Thilak, V. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models, 2025. URL https://arxiv.org/abs/2501.12370.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Clark, A., de las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., van den Driessche, G., Rutherford, E., Hennigan, T., Johnson, M., Millican, K., Cassirer, A., Jones, C., Buchatskaya, E., Budden, D., Sifre, L., Osindero, S., Vinyals, O., Rae, J., Elsen, E., Kavukcuoglu, K., and Simonyan, K. Unified scaling laws for routed language models, 2022.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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, 2024.

DeepSeek-AI, Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Yang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Chen, J., Yuan, J., Qiu, J., Song, J., Dong, K., Gao, K., Guan, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Pan, R., Xu, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Zheng, S., Wang, T., Pei, T., Yuan, T., Sun, T., Xiao, W. L., Zeng, W., An, W., Liu, W., Liang, W., Gao, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Chen, X., Nie, X., Sun, X., Wang, X., Liu, X., Xie, X., Yu, X., Song, X., Zhou, X., Yang, X., Lu, X., Su, X., Wu, Y., Li, Y. K., Wei, Y. X., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Zheng, Y., Zhang, Y., Xiong, Y., Zhao, Y., He, Y., Tang, Y., Piao, Y., Dong, Y., Tan, Y., Liu, Y., Wang, Y., Guo, Y., Zhu, Y., Wang, Y., Zou, Y., Zha, Y., Ma, Y., Yan, Y., You, Y., Liu, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Huang, Z., Zhang, Z., Xie, Z., Hao, Z., Shao, Z., Wen, Z., Xu, Z., Zhang, Z., Li, Z., Wang, Z., Gu, Z., Li, Z., and Xie, Z. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024a. URL https://arxiv.org/abs/2405.04434.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y.,

Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. Deepseek-v3 technical report, 2024b.

Doubov, S., Sardana, N., and Chiley, V. Sparse upcycling: Inference inefficient finetuning, 2024. URL https://arxiv.org/abs/2411.08968.

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., Zhou, Z., Wang, T., Wang, Y. E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q. V., Wu, Y., Chen, Z., and Cui, C. Glam: Efficient scaling of language models with mixture-of-experts, 2022.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Damlaj, I., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N.,

Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Albiero, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., and Jiang, L. Llmcarbon: Modeling the end-to-end carbon footprint of large language models, 2024.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022.

Gale, T., Narayanan, D., Young, C., and Zaharia, M. Megablocks: Efficient sparse training with mixture-of-experts, 2022. URL https://arxiv.org/abs/2211.15841.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., 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., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Krajewski, J., Ludziejewski, J., Adamczewski, K., Pióro, M., Krutul, M., Antoniak, S., Ciebiera, K., Król, K., Odrzygóźdź, T., Sankowski, P., Cygan, M., and Jaszczur, S. Scaling laws for fine-grained mixture of experts, 2024. URL https://arxiv.org/abs/2402.07871.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. Race: Large-scale reading comprehension dataset from examinations, 2017. URL https://arxiv.org/abs/1704.04683.

Lewis, M., Bhosale, S., Dettmers, T., Goyal, N., and Zettlemoyer, L. Base layers: Simplifying training of large, sparse models, 2021.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Ludziejewski, J., Pióro, M., Krajewski, J., Stefaniak, M., Krutul, M., Małaśnicki, J., Cygan, M., Sankowski, P., Adamczewski, K., Miłoś, P., and Jaszczur, S. Joint moe scaling laws: Mixture of experts can be memory efficient, 2025. URL https://arxiv.org/abs/2502.05172.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

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., Farhadi, A., Smith, N. A., Koh, P. W., Singh, A., and Hajishirzi, H. Olmoe: Open mixture-of-experts language models, 2024. URL https://arxiv.org/abs/2409.02060.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson,

C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Parmar, J., Prabhumoye, S., Jennings, J., Patwary, M., Subramanian, S., Su, D., Zhu, C., Narayanan, D., Jhunjhunwala, A., Dattagupta, A., Jawa, V., Liu, J., Mahabaleshwarkar, A., Nitski, O., Brundyn, A., Maki, J., Martinez, M., You, J., Kamalu, J., LeGresley, P., Fridman, D., Casper, J., Aithal, A., Kuchaiev, O., Shoeybi, M., Cohen, J., and Catanzaro, B. Nemotron-4 15b technical report, 2024a. URL https://arxiv.org/abs/2402.16819.

Parmar, J., Satheesh, S., Patwary, M., Shoeybi, M., and Catanzaro, B. Reuse, don't retrain: A recipe for continued pretraining of language models, 2024b. URL https://arxiv.org/abs/2407.07263.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Sap, M., Rashkin, H., Chen, D., LeBras, R., and Choi, Y. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

Shazeer, N. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

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

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Tan, S., Shen, Y., Panda, R., and Courville, A. Scattered mixture-of-experts implementation, 2024. URL https://arxiv.org/abs/2403.08245.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran,

D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Güra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Ágoston Weisz, Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M. A., Baeuml, M., Chen, Z., Shafey, L. E., Zhang, Y., Sercinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rrustemi, A., Clay, N., Crone, P., Kocisky, T., Zhao, J., Perz, B., Yu, D., Howard, H., Bloniarz, A., Rae, J. W., Lu, H., Sifre, L., Maggioni, M., Alcober, F., Garrette, D., Barnes, M., Thakoor, S., Austin, J., Barth-Maron, G., Wong, W., Joshi, R., Chaabouni, R., Fatiha, D., Ahuja, A., Tomar, G. S., Senter, E., Chadwick, M., Kornakov, I., Attaluri, N., Iturrate, I., Liu, R., Li, Y., Cogan, S., Chen, J., Jia, C., Gu, C., Zhang, Q., Grimstad, J., Hartman, A. J., Garcia, X., Pillai, T. S., Devlin, J., Laskin, M., de Las Casas, D., Valter, D., Tao, C., Blanco, L., Badia, A. P., Reitter, D., Chen, M., Brennan, J., Rivera, C., Brin, S., Iqbal, S., Surita, G., Labanowski, J., Rao, A., Winkler, S., Parisotto, E., Gu, Y., Olszewska, K., Addanki, R., Miech, A., Louis, A., Teplyashin, D., Brown, G., Catt, E., Balaguer, J., Xiang, J., Wang, P., Ashwood, Z., Briukhov, A., Webson, A., Ganapathy, S., Sanghavi, S., Kannan, A., Chang, M.-W., Stjerngren, A., Djolonga, J., Sun, Y., Bapna, A., Aitchison, M., Pejman, P., Michalewski, H., Yu, T., Wang, C., Love, J., Ahn, J., Bloxwich, D., Han, K., Humphreys, P., Sellam, T., Bradbury, J., Godbole, V., Samangooei, S., Damoc, B., Kaskasoli, A., Arnold, S. M. R., Vasudevan, V., Agrawal, S., Riesa, J., Lepikhin, D., Tanburn, R., Srinivasan, S., Lim, H., Hodkinson, S., Shyam, P., Ferret, J., Hand, S., Garg, A., Paine, T. L., Li, J., Li, Y., Giang, M., Neitz, A., Abbas, Z., York, S., Reid, M., Cole, E., Chowdhery, A., Das, D., Rogozińska, D., Nikolaev, V., Sprechmann, P., Nado, Z., Zilka, L., Prost, F., He, L., Monteiro, M., Mishra, G., Welty, C., Newlan, J., Jia, D., Allamanis, M., Hu, C. H., de Liedekerke, R., Gilmer, J., Saroufim, C., Rijhwani, S., Hou, S., Shrivastava, D., Baddepudi, A., Goldin, A., Ozturel, A., Cassirer, A., Xu, Y., Sohn, D., Sachan, D., Amplayo, R. K., Swanson, C., Petrova, D., Narayan, S., Guez, A., Brahma, S., Landon, J., Patel, M., Zhao, R., Villela, K., Wang, L., Jia, W., Rahtz, M., Giménez, M., Yeung, L., Keeling, J., Georgiev, P., Mincu, D., Wu, B., Haykal, S., Saputro, R., Vodrahalli, K., Qin, J., Cankara, Z., Sharma, A., Fernando, N., Hawkins, W., Neyshabur, B., Kim, S., Hutter, A., Agrawal, P., Castro-Ros, A., van den Driessche, G., Wang, T., Yang, F., yiin Chang, S., Komarek, P., McIlroy, R., Lučić, M., Zhang, G., Farhan, W., Sharman, M., Natsev, P., Michel, P., Bansal, Y., Qiao, S., Cao, K., Shakeri, S., Butterfield, C., Chung, J., Rubenstein, P. K., Agrawal, S., Mensch, A., Soparkar, K., Lenc, K., Chung, T., Pope, A., Maggiore, L., Kay, J., Jhakra, P., Wang, S., Maynez, J., Phuong, M., Tobin, T., Tacchetti, A., Trebacz, M., Robinson, K., Katariya, Y., Riedel, S., Bailey, P., Xiao, K., Ghelani, N., Aroyo, L., Slone, A., Houlsby, N., Xiong, X., Yang, Z., Gribovskaya, E., Adler, J., Wirth, M., Lee, L., Li, M., Kagohara, T., Pavagadhi, J., Bridgers, S., Bortsova, A., Ghemawat, S., Ahmed, Z., Liu, T., Powell, R., Bolina, V., Iinuma, M., Zablotskaia, P., Besley, J., Chung, D.-W., Dozat, T., Comanescu, R., Si, X., Greer, J., Su, G., Polacek, M., Kaufman, R. L., Tokumine, S., Hu, H., Buchatskaya, E., Miao, Y., Elhawaty, M., Siddhant, A., Tomasev, N., Xing, J., Greer, C., Miller, H., Ashraf, S., Roy, A., Zhang, Z., Ma, A., Filos, A., Besta, M., Blevins, R., Klimenko, T., Yeh, C.-K., Changpinyo, S., Mu, J., Chang, O., Pajarskas, M., Muir, C., Cohen, V., Lan, C. L., Haridasan, K., Marathe, A., Hansen, S., Douglas, S., Samuel, R., Wang, M., Austin, S., Lan, C., Jiang, J., Chiu, J., Lorenzo, J. A., Sjösund, L. L., Cevey, S., Gleicher, Z., Avrahami, T., Boral, A., Srinivasan, H., Selo, V., May, R., Aisopos, K., Hussenot, L., Soares, L. B., Baumli, K., Chang, M. B., Recasens, A., Caine, B., Pritzel, A., Pavetic, F., Pardo, F., Gergely, A., Frye, J., Ramasesh, V., Horgan, D., Badola, K., Kassner, N., Roy, S., Dyer, E., Campos, V. C., Tomala, A., Tang, Y., Badawy, D. E., White, E., Mustafa, B., Lang, O., Jindal, A., Vikram, S., Gong, Z., Caelles, S., Hemsley, R., Thornton, G., Feng, F., Stokowiec, W., Zheng, C., Thacker, P., Çağlar Ünlü, Zhang, Z., Saleh, M., Svensson, J., Bileschi, M., Patil, P., Anand, A., Ring, R., Tsihlas, K., Vezer, A., Selvi, M., Shevlane, T., Rodriguez, M., Kwiatkowski, T., Daruki, S., Rong, K., Dafoe, A., FitzGerald, N., Gu-Lemberg, K., Khan, M., Hendricks, L. A., Pellat, M., Feinberg, V., Cobon-Kerr, J., Sainath, T., Rauh, M., Hashemi, S. H., Ives, R., Hasson, Y., Noland, E., Cao, Y., Byrd, N., Hou, L., Wang, Q., Sottiaux, T., Paganini, M., Lespiau, J.-B., Moufarek, A., Hassan, S., Shivakumar, K., van Amersfoort, J., Mandhane, A., Joshi, P., Goyal, A., Tung, M., Brock, A., Sheahan, H., Misra, V., Li, C., Rakićević, N., Dehghani, M., Liu, F., Mittal, S., Oh, J., Noury, S., Sezener, E., Huot, F., Lamm, M., Cao, N. D., Chen, C., Mudgal, S., Stella, R., Brooks, K., Vasudevan, G., Liu, C., Chain, M., Melinkeri, N., Cohen, A., Wang, V., Seymore, K., Zubkov, S., Goel, R., Yue, S., Krishnakumaran, S., Albert, B., Hurley, N., Sano, M., Mohananey, A., Joughin, J., Filonov, E., Kepa, T., Eldawy, Y., Lim, J., Rishi, R., Badiezadegan, S., Bos, T., Chang, J., Jain, S., Padmanabhan, S. G. S., Puttagunta, S., Krishna, K., Baker, L., Kalb, N., Bedapudi, V., Kurzrok, A., Lei, S., Yu, A., Litvin, O., Zhou, X., Wu, Z., Sobell, S., Siciliano, A., Papir, A., Neale, R., Bragagnolo, J., Toor, T., Chen, T., Anklin, V., Wang, F., Feng, R., Gholami, M., Ling, K., Liu, L., Walter, J., Moghaddam, H., Kishore, A., Adamek, J., Mercado, T., Mallinson, J., Wandekar, S., Cagle, S.,

Ofek, E., Garrido, G., Lombriser, C., Mukha, M., Sun, B., Mohammad, H. R., Matak, J., Qian, Y., Peswani, V., Janus, P., Yuan, Q., Schelin, L., David, O., Garg, A., He, Y., Duzhyi, O., Älgmyr, A., Lottaz, T., Li, Q., Yadav, V., Xu, L., Chinien, A., Shivanna, R., Chuklin, A., Li, J., Spadine, C., Wolfe, T., Mohamed, K., Das, S., Dai, Z., He, K., von Dincklage, D., Upadhyay, S., Maurya, A., Chi, L., Krause, S., Salama, K., Rabinovitch, P. G., M, P. K. R., Selvan, A., Dektiarev, M., Ghiasi, G., Guven, E., Gupta, H., Liu, B., Sharma, D., Shtacher, I. H., Paul, S., Akerlund, O., Aubet, F.-X., Huang, T., Zhu, C., Zhu, E., Teixeira, E., Fritze, M., Bertolini, F., Marinescu, L.-E., Bölle, M., Paulus, D., Gupta, K., Latkar, T., Chang, M., Sanders, J., Wilson, R., Wu, X., Tan, Y.-X., Thiet, L. N., Doshi, T., Lall, S., Mishra, S., Chen, W., Luong, T., Benjamin, S., Lee, J., Andrejczuk, E., Rabiej, D., Ranjan, V., Styrc, K., Yin, P., Simon, J., Harriott, M. R., Bansal, M., Robsky, A., Bacon, G., Greene, D., Mirylenka, D., Zhou, C., Sarvana, O., Goyal, A., Andermatt, S., Siegler, P., Horn, B., Israel, A., Pongetti, F., Chen, C.-W. L., Selvatici, M., Silva, P., Wang, K., Tolins, J., Guu, K., Yogev, R., Cai, X., Agostini, A., Shah, M., Nguyen, H., Donnaile, N. , Pereira, S., Friso, L., Stambler, A., Kurzrok, A., Kuang, C., Romanikhin, Y., Geller, M., Yan, Z., Jang, K., Lee, C.-C., Fica, W., Malmi, E., Tan, Q., Banica, D., Balle, D., Pham, R., Huang, Y., Avram, D., Shi, H., Singh, J., Hidey, C., Ahuja, N., Saxena, P., Dooley, D., Potharaju, S. P., O'Neill, E., Gokulchandran, A., Foley, R., Zhao, K., Dusenberry, M., Liu, Y., Mehta, P., Kotikalapudi, R., Safranek-Shrader, C., Goodman, A., Kessinger, J., Globen, E., Kolhar, P., Gorgolewski, C., Ibrahim, A., Song, Y., Eichenbaum, A., Brovelli, T., Potluri, S., Lahoti, P., Baetu, C., Ghorbani, A., Chen, C., Crawford, A., Pal, S., Sridhar, M., Gurita, P., Mujika, A., Petrovski, I., Cedoz, P.-L., Li, C., Chen, S., Santo, N. D., Goyal, S., Punjabi, J., Kappaganthu, K., Kwak, C., LV, P., Velury, S., Choudhury, H., Hall, J., Shah, P., Figueira, R., Thomas, M., Lu, M., Zhou, T., Kumar, C., Jurdi, T., Chikkerur, S., Ma, Y., Yu, A., Kwak, S., Ähdel, V., Rajayogam, S., Choma, T., Liu, F., Barua, A., Ji, C., Park, J. H., Hellendoorn, V., Bailey, A., Bilal, T., Zhou, H., Khatir, M., Sutton, C., Rzadkowski, W., Macintosh, F., Shagin, K., Medina, P., Liang, C., Zhou, J., Shah, P., Bi, Y., Dankovics, A., Banga, S., Lehmann, S., Bredesen, M., Lin, Z., Hoffmann, J. E., Lai, J., Chung, R., Yang, K., Balani, N., Bražinskas, A., Sozanschi, A., Hayes, M., Alcalde, H. F., Makarov, P., Chen, W., Stella, A., Snijders, L., Mandl, M., Kärrman, A., Nowak, P., Wu, X., Dyck, A., Vaidyanathan, K., R, R., Mallet, J., Rudominer, M., Johnston, E., Mittal, S., Udathu, A., Christensen, J., Verma, V., Irving, Z., Santucci, A., Elsayed, G., Davoodi, E., Georgiev, M., Tenney, I., Hua, N., Cideron, G., Leurent, E., Alnahlawi, M., Georgescu, I., Wei, N., Zheng, I., Scandinaro, D., Jiang, H., Snoek, J., Sundararajan, M., Wang, X., Ontiveros, Z., Karo, I., Cole, J., Rajashekhar, V., Tumeh, L., Ben-David, E., Jain, R., Uesato, J., Datta, R., Bunyan, O., Wu, S., Zhang, J., Stanczyk, P., Zhang, Y., Steiner, D., Naskar, S., Azzam, M., Johnson, M., Paszke, A., Chiu, C.-C., Elias, J. S., Mohiuddin, A., Muhammad, F., Miao, J., Lee, A., Vieillard, N., Park, J., Zhang, J., Stanway, J., Garmon, D., Karmarkar, A., Dong, Z., Lee, J., Kumar, A., Zhou, L., Evens, J., Isaac, W., Irving, G., Loper, E., Fink, M., Arkatkar, I., Chen, N., Shafran, I., Petrychenko, I., Chen, Z., Jia, J., Levskaya, A., Zhu, Z., Grabowski, P., Mao, Y., Magni, A., Yao, K., Snaider, J., Casagrande, N., Palmer, E., Suganthan, P., Castaño, A., Giannoumis, I., Kim, W., Rybiński, M., Sreevatsa, A., Prendki, J., Soergel, D., Goedeckemeyer, A., Gierke, W., Jafari, M., Gaba, M., Wiesner, J., Wright, D. G., Wei, Y., Vashisht, H., Kulizhskaya, Y., Hoover, J., Le, M., Li, L., Iwuanyanwu, C., Liu, L., Ramirez, K., Khorlin, A., Cui, A., LIN, T., Wu, M., Aguilar, R., Pallo, K., Chakladar, A., Perng, G., Abellan, E. A., Zhang, M., Dasgupta, I., Kushman, N., Penchev, I., Repina, A., Wu, X., van der Weide, T., Ponnapalli, P., Kaplan, C., Simsa, J., Li, S., Dousse, O., Yang, F., Piper, J., Ie, N., Pasumarthi, R., Lintz, N., Vijayakumar, A., Andor, D., Valenzuela, P., Lui, M., Paduraru, C., Peng, D., Lee, K., Zhang, S., Greene, S., Nguyen, D. D., Kurylowicz, P., Hardin, C., Dixon, L., Janzer, L., Choo, K., Feng, Z., Zhang, B., Singhal, A., Du, D., McKinnon, D., Antropova, N., Bolukbasi, T., Keller, O., Reid, D., Finchelstein, D., Raad, M. A., Crocker, R., Hawkins, P., Dadashi, R., Gaffney, C., Franko, K., Bulanova, A., Leblond, R., Chung, S., Askham, H., Cobo, L. C., Xu, K., Fischer, F., Xu, J., Sorokin, C., Alberti, C., Lin, C.-C., Evans, C., Dimitriev, A., Forbes, H., Banarse, D., Tung, Z., Omernick, M., Bishop, C., Sterneck, R., Jain, R., Xia, J., Amid, E., Piccinno, F., Wang, X., Banzal, P., Mankowitz, D. J., Polozov, A., Krakovna, V., Brown, S., Bateni, M., Duan, D., Firoiu, V., Thotakuri, M., Natan, T., Geist, M., tan Girgin, S., Li, H., Ye, J., Roval, O., Tojo, R., Kwong, M., Lee-Thorp, J., Yew, C., Sinopalnikov, D., Ramos, S., Mellor, J., Sharma, A., Wu, K., Miller, D., Sonnerat, N., Vnukov, D., Greig, R., Beattie, J., Caveness, E., Bai, L., Eisenschlos, J., Korchemniy, A., Tsai, T., Jasarevic, M., Kong, W., Dao, P., Zheng, Z., Liu, F., Yang, F., Zhu, R., Teh, T. H., Sanmiya, J., Gladchenko, E., Trdin, N., Toyama, D., Rosen, E., Tavakkol, S., Xue, L., Elkind, C., Woodman, O., Carpenter, J., Papamakarios, G., Kemp, R., Kafle, S., Grunina, T., Sinha, R., Talbert, A., Wu, D., Owusu-Afriyie, D., Du, C., Thornton, C., Pont-Tuset, J., Narayana, P., Li, J., Fatehi, S., Wieting, J., Ajmeri, O., Uria, B., Ko, Y., Knight, L., Héliou, A., Niu, N., Gu, S., Pang, C., Li, Y., Levine, N., Stolovich, A., Santamaria-Fernandez, R., Goenka, S., Yustalim, W., Strudel, R., Elqursh, A., Deck, C., Lee, H., Li, Z., Levin, K., Hoffmann, R., Holtmann-Rice, D., Bachem, O., Arora, S., Koh, C., Yeganeh, S. H., Põder, S.,

Tariq, M., Sun, Y., Ionita, L., Seyedhosseini, M., Tafti, P., Liu, Z., Gulati, A., Liu, J., Ye, X., Chrzaszcz, B., Wang, L., Sethi, N., Li, T., Brown, B., Singh, S., Fan, W., Parisi, A., Stanton, J., Koverkathu, V., Choquette-Choo, C. A., Li, Y., Lu, T., Ittycheriah, A., Shroff, P., Varadarajan, M., Bahargam, S., Willoughby, R., Gaddy, D., Desjardins, G., Cornero, M., Robenek, B., Mittal, B., Albrecht, B., Shenoy, A., Moiseev, F., Jacobsson, H., Ghaffarkhah, A., Rivière, M., Walton, A., Crepy, C., Parrish, A., Zhou, Z., Farabet, C., Radebaugh, C., Srinivasan, P., van der Salm, C., Fidjeland, A., Scellato, S., Latorre-Chimoto, E., Klimczak-Plucińska, H., Bridson, D., de Cesare, D., Hudson, T., Mendolicchio, P., Walker, L., Morris, A., Mauger, M., Guseynov, A., Reid, A., Odoom, S., Loher, L., Cotruta, V., Yenugula, M., Grewe, D., Petrushkina, A., Duerig, T., Sanchez, A., Yadlowsky, S., Shen, A., Globerson, A., Webb, L., Dua, S., Li, D., Bhupatiraju, S., Hurt, D., Qureshi, H., Agarwal, A., Shani, T., Eyal, M., Khare, A., Belle, S. R., Wang, L., Tekur, C., Kale, M. S., Wei, J., Sang, R., Saeta, B., Liechty, T., Sun, Y., Zhao, Y., Lee, S., Nayak, P., Fritz, D., Vuyyuru, M. R., Aslanides, J., Vyas, N., Wicke, M., Ma, X., Eltyshev, E., Martin, N., Cate, H., Manyika, J., Amiri, K., Kim, Y., Xiong, X., Kang, K., Luisier, F., Tripuraneni, N., Madras, D., Guo, M., Waters, A., Wang, O., Ainslie, J., Baldridge, J., Zhang, H., Pruthi, G., Bauer, J., Yang, F., Mansour, R., Gelman, J., Xu, Y., Polovets, G., Liu, J., Cai, H., Chen, W., Sheng, X., Xue, E., Ozair, S., Angermueller, C., Li, X., Sinha, A., Wang, W., Wiesinger, J., Koukoumidis, E., Tian, Y., Iyer, A., Gurumurthy, M., Goldenson, M., Shah, P., Blake, M., Yu, H., Urbanowicz, A., Palomaki, J., Fernando, C., Durden, K., Mehta, H., Momchev, N., Rahimtoroghi, E., Georgaki, M., Raul, A., Ruder, S., Redshaw, M., Lee, J., Zhou, D., Jalan, K., Li, D., Hechtman, B., Schuh, P., Nasr, M., Milan, K., Mikulik, V., Franco, J., Green, T., Nguyen, N., Kelley, J., Mahendru, A., Hu, A., Howland, J., Vargas, B., Hui, J., Bansal, K., Rao, V., Ghiya, R., Wang, E., Ye, K., Sarr, J. M., Preston, M. M., Elish, M., Li, S., Kaku, A., Gupta, J., Pasupat, I., Juan, D.-C., Someswar, M., M., T., Chen, X., Amini, A., Fabrikant, A., Chu, E., Dong, X., Muthal, A., Buthpitiya, S., Jauhari, S., Hua, N., Khandelwal, U., Hitron, A., Ren, J., Rinaldi, L., Drath, S., Dabush, A., Jiang, N.-J., Godhia, H., Sachs, U., Chen, A., Fan, Y., Taitelbaum, H., Noga, H., Dai, Z., Wang, J., Liang, C., Hamer, J., Ferng, C.-S., Elkind, C., Atias, A., Lee, P., Listík, V., Carlen, M., van de Kerkhof, J., Pikus, M., Zaher, K., Müller, P., Zykova, S., Stefanec, R., Gatsko, V., Hirnschall, C., Sethi, A., Xu, X. F., Ahuja, C., Tsai, B., Stefanoiu, A., Feng, B., Dhandhania, K., Katyal, M., Gupta, A., Parulekar, A., Pitta, D., Zhao, J., Bhatia, V., Bhavnani, Y., Alhadlaq, O., Li, X., Danenberg, P., Tu, D., Pine, A., Filippova, V., Ghosh, A., Limonchik, B., Urala, B., Lanka, C. K., Clive, D., Sun, Y., Li, E., Wu, H., Hongtongsak, K., Li, I., Thakkar, K., Omarov, K., Majmundar, K., Alverson, M., Kucharski, M., Patel, M., Jain, M., Zabelin, M., Pelagatti, P., Kohli, R., Kumar, S., Kim, J., Sankar, S., Shah, V., Ramachandruni, L., Zeng, X., Bariach, B., Weidinger, L., Vu, T., Andreev, A., He, A., Hui, K., Kashem, S., Subramanya, A., Hsiao, S., Hassabis, D., Kavukcuoglu, K., Sadovsky, A., Le, Q., Strohman, T., Wu, Y., Petrov, S., Dean, J., and Vinyals, O. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.

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, 2024. URL https://arxiv.org/abs/2402.01739.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A., Chen, Z., Le, Q., and Laudon, J. Mixture-of-experts with expert choice routing, 2022.

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, 2022. URL https://arxiv.org/abs/2202.08906.

A. Training and Architecture Hyperparameters

Model architecture. We use a standard decoder-only Transformer (Vaswani et al., 2023) architecture, replacing each feed-forward block with a Mixture-of-Experts layer. The models use SwiGLU (Shazeer, 2020) activation and rotary position embeddings (RoPE) with the rotary percentage set to 0.5. We use the tokenizer from (Parmar et al., 2024a), with a vocabulary size of 256,000.

Training data. Models are pretrained on random subsets with up to 300B tokens sampled from a large, diverse multilingual corpus containing both text and code (see (Parmar et al., 2024a)). Before benchmark evaluation, we perform additional continued pretraining on a high-quality, filtered dataset containing alignment-style question-answer pairs. This step helps to improve benchmark signal by adapting models toward the evaluation distribution.

Training hyperparameters. We use the AdamW (Loshchilov & Hutter, 2019) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay set to 0.1. Models are trained with batches of 4M tokens and a sequence length of 2048. The learning rate is set to 2×10^{-4} and follows a cosine schedule with linear warmup over the first 1% of training steps. We initialize model weights with a standard deviation of $\sigma = 0.01$ and apply attention dropout with $p = 0.1$. The MoE capacity factor is set to 1.5, with auxiliary loss coefficient 1×10^{-2} and z-loss coefficient 1×10^{-3}. Gradient clipping is used with a threshold of 1.0. For the continued pretraining phase on the high-quality, filtered dataset, we generally follow the procedure in (Parmar et al., 2024b). We resume from the model checkpoints without loading the optimizer states. The learning rate follows a cosine schedule, starting from η_{end}, the final learning rate of the initial phase, and decaying to $0.1\eta_{\text{end}}$. Continued pretraining uses a token budget equal to 10% of the original pretraining.

A.1. Evaluation metrics

We evaluate the MoE variants from two complementary perspectives:

Efficiency gains. To measure training efficiency, we primarily track pretraining loss, a standard proxy for model quality. For visualization, we plot training loss (smoothed using the moving average), while tables report final validation loss.

Downstream accuracy. Beyond pretraining efficiency, we assess task-specific performance using a suite of established benchmarks: ARC-Easy, ARC-Challenge (Clark et al., 2018), CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2019), RACE (Lai et al., 2017), SocialIQA (Sap et al., 2019), TruthfulQA (Lin et al., 2022), WinoGrande (Sakaguchi et al., 2019). We use 5-shot evaluation for MMLU and zero-shot for all the other metrics.

A.2. Architecture Hyperparameters

Table 5. Hyperparameter configurations for 11B models.

NAME	EXPERTS	ROUTER TOP-K	ACTIVE PARAMETERS	d_{MODEL}	d_{EXPERT}	TOTAL PARAMETERS
1xFLOPs-G1	8	1	2.7B	2048	8192	11.1B
1xFLOPs-G8	64	8	2.7B	2048	1024	11.1B
2xFLOPs-G1	8	2	3.9B	2048	8192	11.1B
2xFLOPs-G8	64	16	3.9B	2048	1024	11.1B

Table 6. Hyperparameter configurations for the 56B models.

NAME	EXPERTS	ROUTER TOP-K	ACTIVE PARAMETERS	d_{MODEL}	d_{EXPERT}	TOTAL PARAMETERS
1xFLOPs-G1	8	1	10.7B	4096	16384	55.8B
1xFLOPs-G8	64	8	10.7B	4096	2048	55.8B
2xFLOPs-G1	8	2	17.1B	4096	16384	55.8B
2xFLOPs-G8	64	16	17.1B	4096	2048	55.8B

B. Model Evaluation Results

Table 7. Benchmark evaluation results for the 11B models. Within each pair (Switch-style, Mixtral-style), we mark the model with a better score for each metric.

BENCHMARK	1×FLOPS-G1	1×FLOPS-G8	2×FLOPS-G1	2×FLOPS-G8
ARC-C	29.7	**32.7**	31.4	**32.8**
ARC-E	58.1	**60.1**	60.6	**62.3**
HELLASWAG	53.0	**57.4**	56.8	**58.3**
OPENBOOKQA	32.0	**33.8**	31.6	**33.2**
PIQA	71.9	**73.0**	**75.0**	74.2
SOCIALIQA	41.6	**42.6**	42.6	**44.3**
WINOGRANDE	52.6	**54.4**	**55.7**	55.5
AVERAGE	48.4	**50.6**	50.5	**51.5**
VALID LOSS	2.233	**2.183**	2.168	**2.166**

Table 8. Evaluation results for the 56B models. Within each pair (Switch-style, Mixtral-style), we mark the model with a better score for each metric.

BENCHMARK	1×FLOPS-G1	1×FLOPS-G8	2×FLOPS-G1	2×FLOPS-G8
ARC-C	44.3	**46.4**	45.6	**47.7**
ARC-E	73.2	**76.1**	74.3	**77.6**
COMMONSENSEQA	60.4	**65.3**	67.6	**68.7**
HELLASWAG	75.0	**77.0**	76.1	**78.1**
MMLU	47.5	**50.1**	50.5	**52.1**
OPENBOOKQA	**42.6**	42.0	42.0	**45.0**
PIQA	80.5	**81.0**	79.9	**81.2**
RACE	58.1	**60.7**	**60.8**	59.3
SOCIALIQA	47.4	**47.7**	46.3	**46.9**
TRUTHFULQA	34.2	**35.2**	36.3	**38.8**
WINOGRANDE	66.9	**67.5**	67.9	**70.2**
AVERAGE	57.3	**59.0**	58.8	**60.5**
VALID LOSS	1.811	**1.779**	1.780	**1.757**

C. Limitations and Future Work

In this work, we aim to provide a comprehensive comparison of standard and fine-grained MoE models. However, the scope and focus of this work naturally lead to several avenues for future research. Notably, a key assumption throughout this report is uniform hardware utilization across different MoE architectures, considering only training steps and FLOPs. In practice, the exact implementation and hardware setup can lead to variations in Model FLOPs Utilization (MFU) depending on the model type. This variability could affect the efficiency gains described herein. Implementing efficient training and inference for fine-grained MoE remains an important challenge. We refer to related work (Tan et al., 2024; DeepSeek-AI et al., 2024b; Gale et al., 2022; Doubov et al., 2024) and leave a detailed exploration of these implementation and hardware considerations for future studies. Another important direction for future work involves further scaling the dataset size. Our experiments present comparisons using setups with a ratio of tokens to active parameters between 6 and 28. This range is close to the $\sim$20 tokens/parameter rule of thumb observed for dense models in (Hoffmann et al., 2022). However, many modern LLMs (Touvron et al., 2023b; Dubey et al., 2024; Parmar et al., 2024a) are significantly overtrained to maximize evaluation results and inference efficiency. Therefore, it is important to examine the extent to which our conclusions hold for setups involving overtrained models and trillion-scale datasets. Finally, this work focuses exclusively on the pretraining phase. Unique challenges may emerge during subsequent phases like fine-tuning and post-training, representing another area for future investigation.

D. Distribution of the Router Logits

Figure 3. Distribution of router logits in the initial layer *(top row)* middle layer *(center row)* and final layer *(bottom row)* of the 1xFLOPs-G8 model.

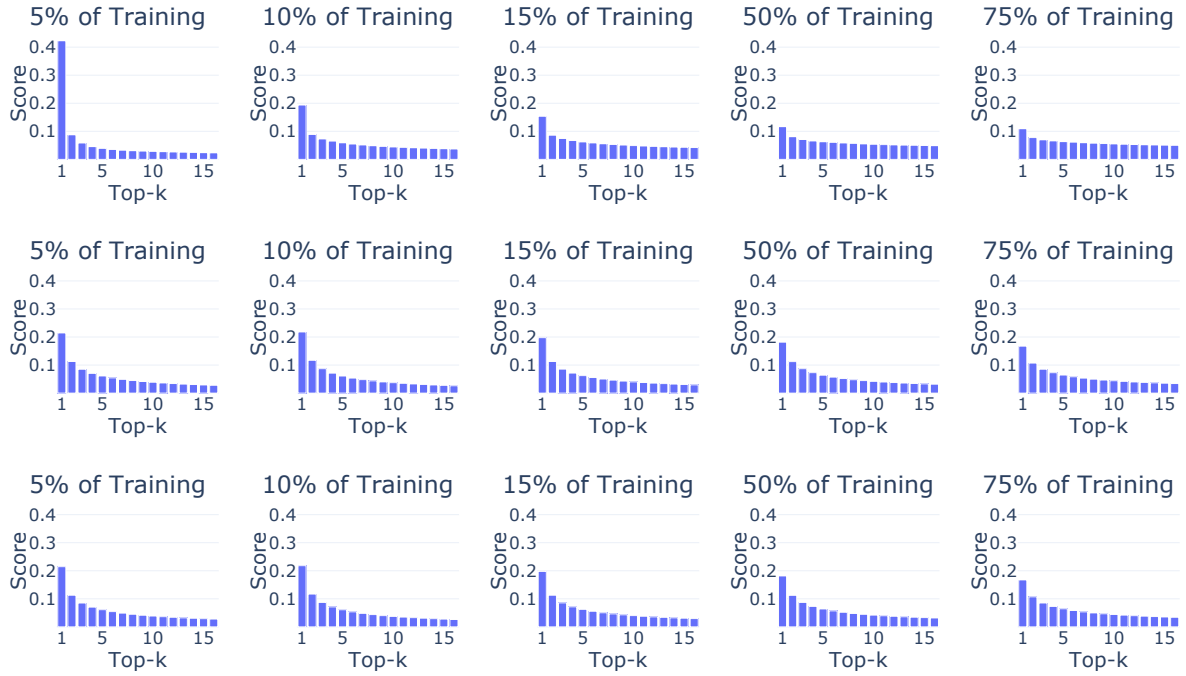


Figure 4. Distribution of router logits in the initial layer *(top row)* middle layer *(center row)* and final layer *(bottom row)* of the 2xFLOPs-G8 model.

E. Analysis of Balance Between Experts

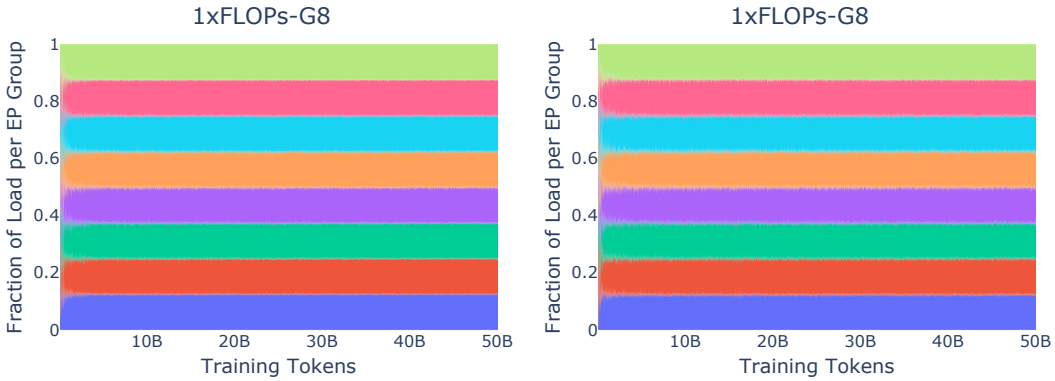


Figure 5. Fraction of load assigned to each Expert Parallel group for the 11B 1xFLOPs-G8 model. *(left)*: Distribution in the middle (12th) layer. *(right)*: Distribution in the final (24th) layer.

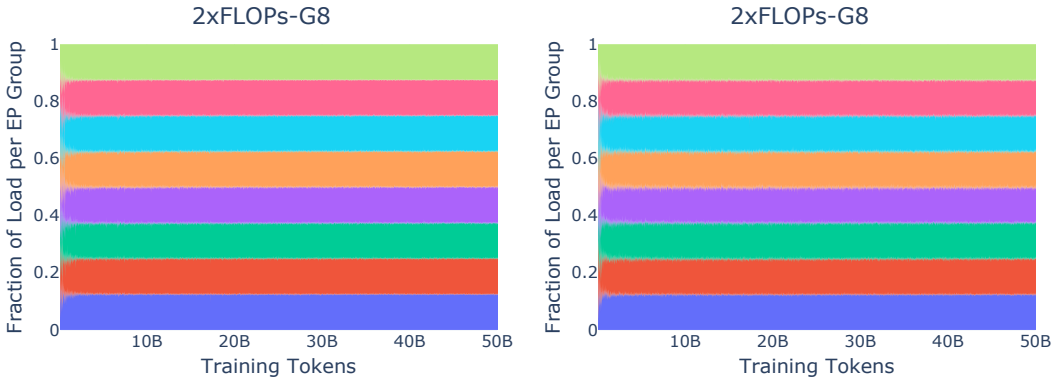


Figure 6. Fraction of load assigned to each Expert Parallel group for the 11B 2xFLOPs-G8 model. *(left)*: Distribution in the middle (12th) layer. *(right)*: Distribution in the final (24th) layer.

F. Related Work

In the context of Transformers, Mixture of Experts is constructed by replacing the Feed-Forward component with a set of *experts*. Each expert is typically a subnetwork of the same shape as the corresponding Feed-Forward layer. Each input x is routed to a subset of experts. The output y of the layer can be defined as

$$y = \Sigma_{i \in \tau} R_i(x) E_i(x),$$

where τ is the set of selected selected Top-k experts for this input and R is the routing network defining the expert scores for the given input. This gating network is a simple linear layer, followed by softmax normalization and Top-k choice.

Mixture of Experts was originally introduced by (Shazeer et al., 2017) and later applied in a series of works in the context of Transformers, including Switch (Fedus et al., 2022), GLaM (Du et al., 2022), and OpenMoE (Xue et al., 2024). Various variations of the routing algorithm have been introduced, including BASE layers (Lewis et al., 2021) and Expert Choice (Zhou et al., 2022). (Clark et al., 2022) presented the first scaling laws for MoE. (Dai et al., 2024) introduced fine-grained MoE architecture, later scaled in a series of models (DeepSeek-AI et al., 2024a;b). (Krajewski et al., 2024) introduced the concept of granularity and derived scaling laws including granularity, training length and model size. (Muennighoff et al., 2024) trained OLMoE, a fully open Mixture of Experts model. (Abnar et al., 2025; Ludziejewski et al., 2025) propose MoE scaling laws including factors like sparsity and total memory usage of the model.