

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 MIXTURE OF HETEROGENEOUS GROUPED EXPERTS FOR LANGUAGE MODELING

Anonymous authors

Paper under double-blind review

## ABSTRACT

Mixture-of-Experts (MoE) offers superior performance over dense models. However, current MoEs impose a critical limitation by enforcing uniform expert sizes, restricting the model’s ability to dynamically match computational resources with token-specific requirements. Despite several attempts on heterogeneous experts have been made, they struggle either with limited performance and inefficient parameter utilization or unbalanced GPU utilization, there is still a lack of general heterogeneous MoE architecture. To this end, we present Mixture of Heterogeneous Grouped Experts (MoHGE), an innovative MoE architecture that introduces a two-level routing mechanism and enables more nuanced and efficient expert selection tailored to each input token’s characteristics. We also propose a Group-Wise Auxiliary Loss to enhance efficient parameter utilization without compromising model performance. To address the resulted workload imbalance challenges, we develop: (1) an All-size Group-decoupling Allocation strategy and (2) Intra-Group Experts Auxiliary Loss, collectively ensuring balanced GPU utilization. Extensive evaluations on multiple benchmarks demonstrate that MoHGE achieves comparable performance to state-of-the-art MoE architectures while reducing total parameter count by approximately 20% and maintaining balanced GPU utilization. Our work establishes a new paradigm for resource-aware MoE design, better aligning computational allocation with actual inference demands.

## 1 INTRODUCTION

Transformer-based large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Bai et al., 2023; Liu et al., 2024a) have achieved remarkable success across a wide range of natural language processing (NLP) tasks. According to scaling laws (Kaplan et al., 2020), larger models consistently deliver better performance, and recent studies (Wei et al., 2022) have shown that scaling can also give rise to emergent abilities. However, the computational cost of training and deploying such large models grows exponentially (Thompson et al., 2020), creating a critical bottleneck for both research and real-world applications.

Mixture-of-Experts (MoE) architectures, originally proposed in Jacobs et al. (1991) and Jordan & Jacobs (1994), offer an effective solution by enabling sparse activation: Only a small subset of the model parameters are engaged in per inference step, allowing the model to scale efficiently without proportionally increasing computational overhead.

Despite this advantage, most existing MoE models consist of experts with identical sizes and structures. This homogeneity poses a limitation when generating tokens of varying difficulty: some tokens are easy to predict, while others require more sophisticated reasoning. To address this, recent approaches such as MoDSE (Sun et al., 2024) and HMoE (Wang et al., 2024) have explored the use of experts with different sizes.

However, MoDSE employs a routing strategy that promotes uniform routing probabilities among experts, which fail to route input tokens to the most suitable experts, leading to inefficient parameter utilization. Since experts have different parameter sizes, this setting limits the combination of experts, making it impossible to select multiple smallest or largest experts, missing opportunities for better efficiency or performance. HMoE mentions the idea of hybrid heterogeneous–homogeneous experts as a promising direction, but does not explicitly explore this design. Moreover, it suffers

054 from significant GPU utilization imbalance due to uneven parameter sizes, ultimately degrading  
 055 training efficiency and limiting its scalability.  
 056

057 In this paper, we first divide experts into multiple groups, where experts within each group share  
 058 identical parameter sizes, while the expert sizes vary across groups. We then introduce a two-level  
 059 routing strategy to deliver more diverse and nuanced expert combinations. We further propose a  
 060 Group-Wise Auxiliary Loss to enable the selection of expert groups with appropriate parameter  
 061 sizes, based on the task difficulty. This ensures more efficient parameter utilization by dynamically  
 062 matching computational resources to token-specific requirements. To address GPU load imbalance,  
 063 we propose an All-size Group-decoupling Allocation strategy, which places an equal number of  
 064 experts from each group onto the each GPU. This strategy guarantees that each GPU has the same  
 065 memory consumption. Further, we propose an Intra-Group Experts Auxiliary Loss to maintain  
 066 balanced routing probabilities within each expert group, ensuring uniform GPU utilization. We  
 067 refer to this novel architecture as the Mixture of Heterogeneous Grouped Experts (MoHGE). Our  
 068 contributions are summarized as follows:  
 069

- **Novel Architecture:** We propose a novel MoE architecture, MoHGE, that achieves precise capacity match based on task difficulty and efficient GPU utilization by incorporating the two-level routing strategy and the Group-Wise Auxiliary Loss.
- **Load Balance:** To ensure balanced GPU utilization, we propose the All-size Group-decoupling Allocation strategy and the Intra-Group Experts Auxiliary Loss. Together, these techniques maintain intra-group utilization equilibrium and achieve uniform GPU workloads, ensuring the model’s scalability.
- **Empirical Validation:** Experimental results demonstrate the framework’s effectiveness: MoHGE achieves an accuracy comparable to that of conventional MoE while reducing total parameters. More noteworthy, detailed routing analysis confirms successful balance of GPU utilization and validates our loss functions’ ability to regulate expert activation patterns.

## 081 2 BACKGROUND: MIXTURE OF EXPERTS

082 An MoE layer typically includes the gating model  $G_1(\cdot) \cdots G_N(\cdot)$ , the expert networks  
 083  $E_1(\cdot) \cdots E_N(\cdot)$ , and the routing mechanism, where  $N$  denotes the number of experts. The gating  
 084 model serves as the mathematical implementation of a router, determining how input data is  
 085 allocated to experts. Specifically, the gating model with learnable weights  $W \in \mathbb{R}^{h_{\text{input}} \times h}$  selects  
 086 the top  $k$  experts and combines the outputs of these top  $k$  experts to produce the output  $y \in \mathbb{R}^h$ ,  
 087 where  $h_{\text{input}}$  is the dimension of input  $x$  and  $h$  is the dimension of the hidden layer. The output of  
 088 an MoE layer can be expressed as,  
 089

$$y = \sum_{i=1}^N G_i(x) E_i(x) \quad (1)$$

$$G_i(x) = \text{Softmax}(\text{top}K(H(x))) \quad (2)$$

$$H(X)_i = (x \cdot W)_i \quad (3)$$

$$\text{Top}K(v, k)_i = \begin{cases} v_i, & v_i \in \text{top}k(v) \\ -\infty, & \text{otherwise} \end{cases} \quad (4)$$

## 099 3 MIXTURE OF HETEROGENEOUS GROUPED EXPERTS

### 100 3.1 GROUP-WISE VARIED SIZE EXPERTS

101 Traditional MoE architectures typically employ a gating network that routes inputs to a uniform  
 102 set of experts, all of which have the same model size. However, as shown by Sun et al. (2024),  
 103 the cognitive challenge of predicting the next token varies significantly across different linguistic  
 104 contexts—mirroring the dynamic processing demands seen in human cognition.  
 105

106 Building on this observation, we introduce a novel heterogeneous expert architecture that organizes  
 107 experts into multi-granularity groups. Formally, we structure the expert set  $\{E_1, E_2, E_3, \dots, E_{N_e}\}$

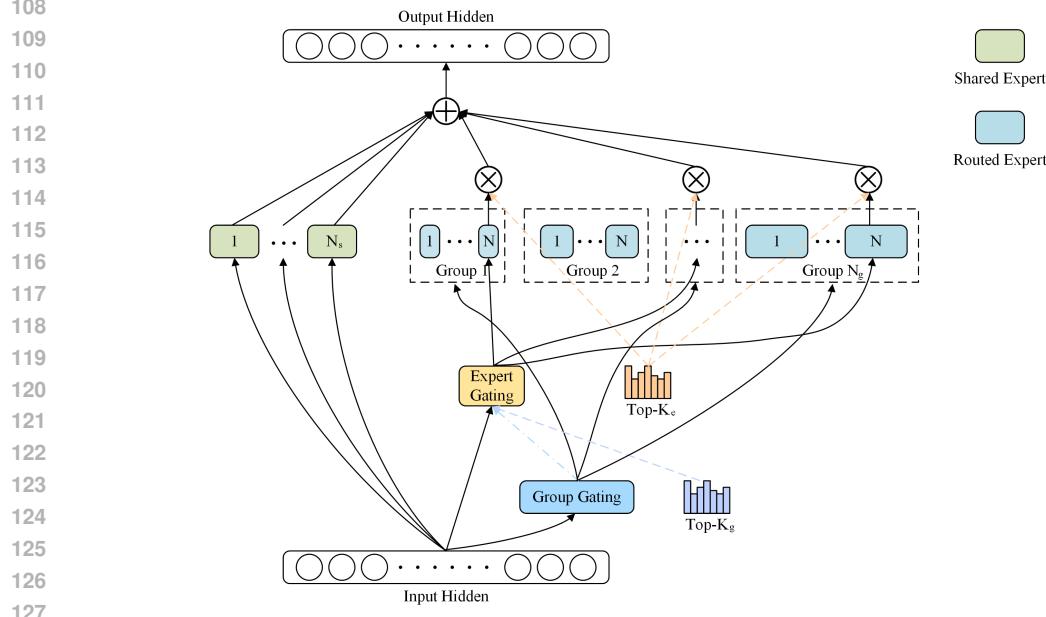


Figure 1: An illustration of our Mixture of Heterogeneous Grouped Experts Layer.

into distinct groups  $\{G_1, G_2, G_3, \dots, G_{N_g}\}$ , where each group contains  $N = N_e/N_g$  experts ( $N_e$  and  $N_g$  denote the total number of experts and groups, respectively). For the convenience of expression, we transform experts from  $E_j$  into  $E_{g,i}$ , where  $g$  represents the group to which the expert belongs and  $i$  represents the index of the expert in the group. Experts within each group share identical parameter sizes, while parameter scales vary across groups according to a predefined progression. Specifically, the hidden dimension of experts in group  $G_i$  is given by:

$$W_i = 2 * W_{\text{base}} - W_{N_g - i} \quad (5)$$

where  $W_{\text{base}}$  represents the base hidden dimension and the  $W_i$  increases as  $i$  increases. This hierarchical organization enables dynamic computation allocation: compact experts efficiently process simpler linguistic patterns, while progressively larger experts with greater capacity handle more complex contextual relationships.

### 3.2 TWO-LEVEL ROUTING MECHANISM

To efficiently manage the hierarchical structure of experts, our two-level routing mechanism operates in two stages. The **group gating model** first selects expert groups based on their relevance to the input, and the **expert gating model** then chooses specific experts within these groups. This staged design ensures that computation is focused on the most relevant experts, reducing overhead by restricting selection to the top- $K_g$  groups.

#### 3.2.1 GROUP GATING MODEL

The group gating model computes scores  $GS$  for all  $N_g$  expert groups. For the  $t$ -th token input  $\mathbf{x}_t$ , the score for the  $g$ -th group is,

$$GS_{g,t} = \text{Sigmoid}(\mathbf{x}_t^T \mathbf{e}_g) \quad (6)$$

where  $\mathbf{e}_g$  is the centroid embedding of the  $g$ -th expert group. The model then selects the  $K_g$  groups with the highest scores, restricting the expert gating model to only route tokens to experts within these groups.

#### 3.2.2 EXPERT GATING MODEL

The expert gating model operates in three phases: **Intra-Group Expert Scores Calculation**, **Experts for Global Selection** and **Global Normalization**.

162 **1. Intra-Group Expert Scores Calculation.** For each selected group, the model computes unnor-  
 163 malized scores for its experts using a group-wise Softmax:  
 164

$$165 \quad ES'_{g,i,t} = \begin{cases} \text{Softmax}(\mathbf{x}_t^T \mathbf{e}_{g,i}), & \text{if } GS_{g,t} \in \text{top}K_g(GS_t) \\ 166 \quad 0, & \text{otherwise} \end{cases} \quad (7)$$

167 where  $e_{g,i}$  is the embedding of the  $i$ -th expert in group  $g$ .  
 168

169 **2. Experts for Global Selection.** The intra-group expert scores are scaled by the group scores to  
 170 reflect group importance:  
 171

$$172 \quad ES''_{g,i,t} = (ES' \cdot GS)_{g,i,t} \quad (8)$$

173 Next, the model selects the top- $K_e$  experts globally. Scores for all other experts are set to zero:  
 174

$$175 \quad ES'''_{g,i,t} = \begin{cases} ES''_{g,i,t}, & \text{if } ES''_{g,i,t} \in \text{top}K_e(ES''_{g,i,t}) \\ 176 \quad 0, & \text{otherwise} \end{cases} \quad (9)$$

177 **3. Global Normalization.** Finally, the selected expert scores are normalized to sum to one:  
 178

$$179 \quad ES_{g,i,t} = \frac{ES'''_{g,i,t}}{\sum_j^{N_g} \sum_k^N ES'''_{j,k,t}} \quad (10)$$

182 This three-step gating strategy enables fine-grained, efficient expert selection by prioritizing both  
 183 group relevance and individual expert utility.  
 184

### 185 3.3 OUTPUT OF MOHGE

187 The output of MoHGE layer is similar to the MoE layer, the outputs of all selected experts are  
 188 multiplied by their corresponding scores and then added together to obtain the final output:  
 189

$$190 \quad y = \sum_{g=1}^{N_g} \sum_{i=1}^N ES_{g,i,t} \cdot E_{g,i}(x_t) \quad (11)$$

### 193 3.4 EFFICIENT PARAMETER UTILIZATION

195 Without regularization, experts with larger parameter sizes tend to dominate the routing decisions  
 196 due to their stronger representational capacity. This dominance can result in inefficient expert usage,  
 197 as smaller expert groups with fewer parameters may not be fully utilized. To address this issue and  
 198 improve parameter utilization, we introduce a slight penalty for expert groups with larger parameter  
 199 sizes. Specifically, we propose **Group-Wise Auxiliary Loss**  $L_G$ , which slightly penalizes expert  
 200 groups with larger parameter sizes.

201 This loss encourages the gating model to consider groups with fewer parameters, leading to more  
 202 efficient parameter utilization. The model ultimately learns to trade off between minimizing cross-  
 203 entropy and reducing parameter-related costs. The group-wise loss is formulated as:  
 204

$$206 \quad L_G = \alpha_G \sum_{i=1}^{N_g} \frac{W_i}{W_{max}} f_i^G p_i^G \quad (12)$$

$$208 \quad f_i^{Grp} = \frac{N_g}{K_g} \sum_{t=1}^T \mathbb{1}(GS_{i,t} \in \text{top}k(GS_t)) \quad (13)$$

$$210 \quad p_i^G = \frac{1}{T} \sum_i^T s_{i,t}^G \quad (14)$$

$$212 \quad s_{i,t}^G = \frac{GS_{i,t}}{\sum_j^{N_g} GS_{j,t}} \quad (15)$$

215 where  $W_i$  is the parameter count of group  $i$ ,  $f_i^G$  is the group's routing frequency, the balance factor  
 $\alpha_G$  is assigned an extremely small value and  $p_i^G$  is its average normalized routing score.

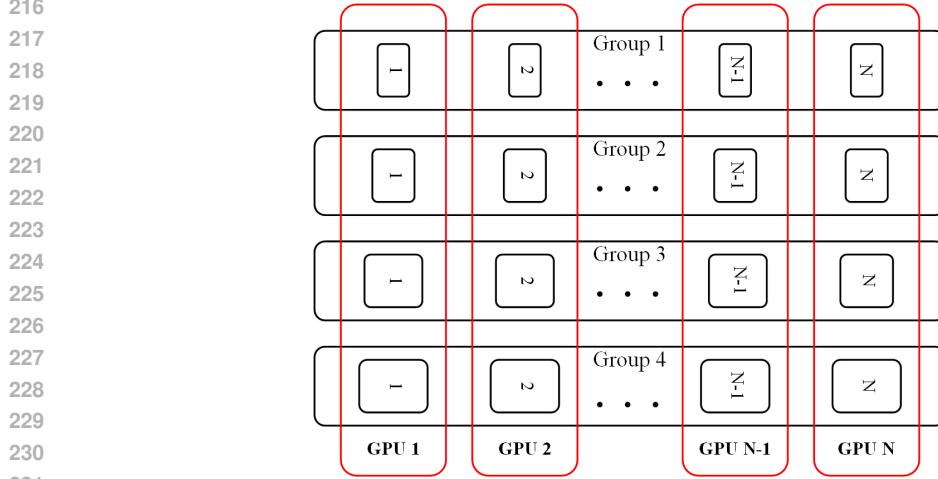


Figure 2: An example of All-size Group-decoupling Allocation.

### 3.5 LOAD BALANCE CONSIDERATION

Experts with larger hidden dimensions (i.e., those exceeding a base width  $W_{\text{base}}$ ) introduce disproportionately higher memory and computational costs. If not carefully managed, this imbalance can lead to severe GPU load imbalances, where certain GPUs become bottlenecks while others remain underutilized. This inefficiency hampers overall training performance and scalability. To mitigate this issue, we introduce **All-size Group-decoupling Allocation** and **Intra-Group Experts Auxiliary Loss**, which work synergistically to achieve a uniform distribution of computational load across GPUs, thus ensuring balanced resource utilization.

#### 3.5.1 ALLOCATION STRATEGY

An All-size expert set consists of the  $i$ -th expert from all groups. Each GPU is assigned multiple such sets, ensuring that the total number of expert parameters on each GPU remains consistent and smoothing out the variance in parameter size across the system. If expert workloads are evenly balanced within each group (which is encouraged by our auxiliary loss design), this approach leads to balanced GPU utilization overall.

As illustrated in Fig. 2 (with  $N_g = 4$ ), each GPU hosts one All-size expert set (e.g., experts  $E_{1,i}, E_{2,i}, E_{3,i}, E_{4,i}$ ). Regardless of the group selection during routing, as long as expert activation within each group is balanced, overall GPU resource usage remains evenly distributed.

#### 3.5.2 INTRA-GROUP EXPERTS AUXILIARY LOSS

In addition to the standard cross-entropy loss, we incorporate an intra-group experts auxiliary loss  $L_E$  adapted from DeepSeekV2 (Liu et al., 2024a) to encourage balanced expert usage during routing. While DeepSeekV2 penalizes imbalance across all experts globally, our approach focuses on experts within each selected group, promoting uniform routing frequencies locally. This design ensures that all experts within an active group are selected with equal frequency during training, leading to better load distribution across GPUs.

The auxiliary loss is defined as:

$$L_E = \alpha_E \sum_{g=1}^{N_g} \sum_{i=1}^N f_{g,i}^E p_{g,i}^E \quad (16)$$

270

271

$$f_{g,i}^E = \frac{N}{K_e} \sum_{t=1}^T \mathbb{1}(ES'_{g,i,t} \in \text{top}K_e(ES'_t)) \quad (17)$$

274

$$p_{g,i}^E = \frac{1}{T} \sum_i^T S_{g,j,t}^{Exp} \quad (18)$$

276

$$S_{g,j,t}^E = \frac{ES'_{g,i,t}}{\sum_j^N ES'_{g,j,t} + \epsilon} \quad (19)$$

278

where  $f_{g,i}^E$  represents the normalized routing frequency of the  $i$ -th expert in group  $g$ ,  $s_{g,i,t}^e$  is the normalized routing score,  $p_{g,i}^E$  is the average selection probability across time steps, the balance factor  $\alpha_E$  is assigned an extremely small value and  $\epsilon$  is a very small constant to ensure that the denominator is not 0.

283

## 284 4 EXPERIMENTS

285

### 286 4.1 EXPERIMENTAL SETUP

288

**Compute Infrastructure.** All models were trained on a 16-node GPU cluster, with each node equipped with eight NVIDIA GPUs. We used the Megatron-LM framework (Shoeybi et al., 2019) to implement our MoHGE variants, as well as the dense and MoE baseline models.

291

**Pretraining Data.** Our pretraining corpus was created by merging and deduplicating three large English datasets: DataComp-LM, FineWeb, and The Pile. The combined corpus underwent standard noise filtering and quality checks to ensure data integrity. For all experiments, we sampled 0.58 trillion tokens from this cleaned, unified corpus.

296

**Model Configurations.** We evaluated three Transformer variants at the 1B 3B, and 14B parameter scales: a Dense model whose parameters are equal to the active parameters of the MoE baseline, a uniform-expert MoE baseline, and our proposed MoHGE architecture with heterogeneous expert groups. The MoE baseline is adapted from DeepSeekV2 (Liu et al., 2024b), with hyperparameters adjusted to align parameter counts across models for fair comparison. Detailed architectural configurations for all evaluated models are summarized in **APPENDIX**.

303

**Training Hyperparameters.** Each MoE model was trained for 2 full epochs on the 0.58 trillion-token corpus, using a fixed sequence length of 4,096. We used the AdamW optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ , and a weight decay of 0.1. A cosine-decay learning rate schedule was applied, starting at  $3 \times 10^{-4}$  and annealing to a minimum of  $3 \times 10^{-5}$ .

308

### 309 4.2 MAIN RESULTS

310

Following OpenCompass protocols (Contributors, 2023), Table 1 reports the zero-shot or few-shot (Kojima et al., 2022; Brown et al., 2020) in-context learning performance of our pretrained MoHGE models on a diverse suite of downstream tasks, including MMLU (Hendrycks et al., 2020), SIQA (Sap et al., 2019), GSM8K (Cobbe et al., 2021), LAMBADA (Paperno et al., 2016), MATH (Hendrycks et al., 2024), PIQA (Bisk et al., 2020) Bisk et al. (2020) and TriviaQA (Joshi et al., 2017).

316

As reported in Table 1, averaged over three evaluate runs, MoHGE consistently outperforms both conventional MoE and dense models across all scales, achieving state-of-the-art results on several benchmarks. Compared to the MoE baseline, MoHGE achieves a more favorable trade-off between parameter efficiency and downstream performance by activating fewer expert parameters while simultaneously requiring fewer total parameters.

322

Specifically, MoHGE reduces the overall parameter count by nearly 20% relative to standard MoE, and the number of activated parameters in the expert layer is reduced by approximately one quarter. This substantial reduction highlights its effectiveness in balancing model capacity with efficiency.

Method	Total Parameters	Activated Parameters of Experts	MMLU	SIQA	GSM8K	LAMBADA	MATH	PIQA	TriviaQA
Dense	0.570B	—	25.41	34.93	1.79	51.87	1.22	44.85	25.05
MoE-1B	1.098B	0.163B	25.38	35.12	1.74	53.20	1.26	46.09	<b>25.86</b>
MoHGE-1B	0.891B	0.122B	<b>25.98</b>	<b>35.17</b>	<b>1.97</b>	<b>53.75</b>	<b>1.29</b>	<b>48.85</b>	25.71
Dense	0.807B	—	26.36	35.30	2.79	61.02	1.33	47.35	34.98
MoE-3B	3.3614B	0.376B	26.22	35.41	3.03	60.86	1.34	<b>49.08</b>	39.16
MoHGE-3B	2.821B	0.295B	<b>26.41</b>	<b>35.56</b>	<b>4.02</b>	<b>62.37</b>	<b>1.36</b>	<b>49.08</b>	<b>39.20</b>
Dense	1.672B	—	30.78	42.29	4.61	68.05	6.81	54.92	50.26
MoE-14B	16.760B	1.191B	31.18	44.28	4.88	67.94	7.29	56.71	51.77
MoHGE-14B	14.122B	0.843B	<b>31.62</b>	<b>45.62</b>	<b>5.73</b>	<b>69.89</b>	<b>9.42</b>	<b>58.73</b>	<b>52.69</b>

Table 1: Comparison between Dense model, MoE baseline and our MoHGE, the highest scores for each benchmark is highlighted in bold. MoHGE achieves slightly better performance while activating the fewest parameters. Furthermore, our model requires fewer total parameters than the baseline in addition to its efficiency advantages.

Benchmark	MoE-1B(hours)	MoHGE-1B(hours)	MoE-3B(hours)	MoHGE-3B(hours)	MoE-14B(hours)	MoHGE-14B(hours)
MMLU	6.90	6.77	9.85	9.58	19.27	18.86
SIQA	0.93	0.90	1.29	1.17	2.51	2.33
GSM8K	0.59	0.62	0.84	0.86	1.63	1.62
LAMBADA	2.24	2.22	3.17	3.03	6.27	6.08
MATH	2.24	2.23	3.18	3.02	6.33	6.09
PIQA	0.85	0.78	1.20	1.09	2.38	2.17
TriviaQA	4.11	3.95	5.78	5.46	11.46	10.85

Table 2: The inference duration of the MoE and MoHGE models on downstream tasks.

The inference times are demonstrated in Table 2. Regarding the slight increase in inference time on GSM8K which is a complex mathematical reasoning task, our routing analysis reveals that the MoHGE tends to select expert groups with larger parameter on GSM8K and this achieves higher accuracy while resulting in more inference time. Altogether, our model achieves relatively faster inference speeds, showing superior inference efficiency.

### 4.3 ABLATION STUDY ON AUXILIARY LOSS COEFFICIENTS

We conduct an ablation study to analyze the effect of different auxiliary loss coefficients on model performance. A coefficient of 0 indicates the absence of the auxiliary loss.

As shown in Table 3, the intra-group experts auxiliary loss yields a modest performance gain and setting  $\alpha_{Exp} = 2.5e - 3$  achieves better results. Combining it with the group-wise auxiliary loss further improves results. Although the group-wise loss contributes only marginally to accuracy, it reduces the number of activated parameters. Based on the trade-off between evaluation performance and computational efficiency, we find that setting  $\alpha_{Exp} = 2.5e - 3$  and  $\alpha_{Grp} = 1e - 4$  enables the our models to achieve an optimal balance.

### 4.4 ANALYSIS ON TOKEN ROUTING

#### 4.4.1 ROUTING ANALYSIS OF LOSS FUNCTION

Building on the optimal configurations identified in Table 3, we conduct experiments on two model configurations:

**Utilizing only intra-group expert auxiliary loss:**  $\alpha_{Exp} = 2.5e - 3$  and  $\alpha_{Grp} = 0$ .

**Combining two loss functions:**  $\alpha_{Exp} = 2.5e - 3$  and  $\alpha_{Grp} = 1e - 4$ .

We statistically analyzed the distribution of 100 million token routes across these configurations. As shown in Fig. 3, the overall route distribution does not exhibit concentration in specific groups under either setup. However, introducing the group routing loss shifts the token routing behavior: instead of predominantly favoring larger expert groups, tokens are distributed toward smaller. This indicates that the group-wise loss encourages the selection of smaller expert groups which can accommodate the current task difficulty in condition of relatively uniform routing distribution.

Model	$\alpha_{Exp}$	$\alpha_{Grp}$	Activated Parameters of Experts	MMLU	SIQA	PIQA	LAMBADA	TriviaQA
MoHGE-1B	0	0	139M	25.43	34.73	47.62	52.20	25.03
	2.5e-3	0	132M	25.61	34.82	47.93	53.35	25.37
	5e-3	0	131M	25.87	34.74	48.77	53.14	25.20
	2.5e-3	1e-4	122M	<b>25.98</b>	<b>35.17</b>	<u>48.85</u>	<b>53.75</b>	<b>25.42</b>
	2.5e-3	1e-3	122M	25.94	<u>35.10</u>	48.28	52.99	25.25
	5e-3	1e-4	119M	<u>25.96</u>	34.86	48.12	53.16	<u>25.39</u>
MoHGE-3B	0	0	324M	25.88	35.29	48.65	61.37	38.01
	2.5e-3	0	307M	26.11	35.45	48.53	61.62	38.53
	5e-3	0	310M	26.03	35.32	48.67	61.75	38.45
	2.5e-3	1e-4	295M	<b>26.31</b>	<b>35.56</b>	<b>49.08</b>	<b>62.37</b>	<b>39.20</b>
	2.5e-3	1e-3	297M	<b>26.36</b>	35.12	48.83	<u>62.10</u>	38.68
	5e-3	1e-4	289M	26.27	<u>35.47</u>	48.21	61.85	38.51
MoHGE-14B	0	0	897M	31.37	44.92	57.94	68.57	51.72
	2.5e-3	0	884M	31.18	45.03	58.07	68.95	51.86
	5e-3	0	875M	<b>31.71</b>	45.39	58.27	68.90	52.29
	2.5e-3	1e-4	843M	<u>31.62</u>	<b>45.62</b>	<b>58.73</b>	<b>69.89</b>	<u>52.49</u>
	2.5e-3	1e-3	854M	30.87	45.07	58.22	69.10	<b>52.75</b>
	5e-3	1e-4	859M	31.38	44.78	58.15	69.85	51.67

Table 3: The evaluation results for varying coefficients of the auxiliary loss function. The highest-performing score for each benchmark is highlighted in bold, while the second-highest score is underlined.

	GPU_1	GPU_2	GPU_3	GPU_4	GPU_5	GPU_6	GPU_7	GPU_8	Avg	Std
Group_1	2.05M	1.89M	1.92M	1.99M	2.02M	1.86M	1.97M	2.01M	1.96M	0.06283
Group_2	1.66M	1.74M	1.82M	1.62M	1.59M	1.74M	1.58M	1.75M	1.69M	0.08166
Group_3	1.58M	1.59M	1.67M	1.50M	1.51M	1.52M	1.64M	1.70M	1.59M	0.07115
Group_4	1.67M	1.58M	1.65M	1.59M	1.67M	1.62M	1.70M	1.51M	1.62M	0.05786
Group_5	1.25M	1.39M	1.38M	1.27M	1.33M	1.38M	1.35M	1.26M	1.33M	0.05452
Group_6	1.45M	1.41M	1.45M	1.31M	1.49M	1.40M	1.50M	1.41M	1.43M	0.05629
Group_7	1.52M	1.60M	1.57M	1.43M	1.41M	1.55M	1.62M	1.43M	1.52M	0.07745
Group_8	1.39M	1.32M	1.46M	1.41M	1.33M	1.35M	1.32M	1.36M	1.37M	0.04630

Table 4: For 14B scale model, the number of tokens routed to each GPU roughly closes to the average value.

#### 4.4.2 ROUTING ANALYSIS OF GPU UTILIZATION

To rigorously evaluate the balancing of GPU utilization, we conduct a GPU-level assessment for 14B scale model by strategically assigning the  $i$ -th expert from each capacity group to the  $i$ -th GPU. This experimental design allows us to precisely track how tokens are distributed across experts of varying sizes on each GPU, which reflects the frequency of token processed by experts of different sizes on each GPU. Table 4 shows that experts of uniform size receive nearly equal routing frequencies across GPUs, indicating balanced intra-group expert and GPU utilization. This confirms that our All-size Group-decoupling Allocation and Intra-group Experts Auxiliary Loss effectively maintain equilibrium in both computational resource loading and expert activation patterns.

## 5 RELATED WORK

The MoE model was originally proposed by Jacobs et al. (1991). Subsequently, Shazeer et al. (2017) introduced Sparsely-Gated Mixture-of-Experts which demonstrate substantial improvements in model capacity and efficiency. Furtherly, SwitchTransformer, proposed by Fedus et al. (2022), incorporated MoE into the Transformer architecture’s Feed-Forward Network layers with simplified MoE routing algorithm, showing great potential in large-scale Transformer models. Typically, MoE models consist of homogeneous experts, each with identical number of parameters, and a predetermined number of experts are activated regardless of the input’s complexity. However, this hinders effective expert specialization and efficient parameter utilization.

Huang et al. (2024) proposed Top-P routing algorithm to address inefficient parameter utilization by assigning different numbers of experts to different tokens. Nevertheless, this method relies on fixed threshold settings and employs a rudimentary approach to difficulty modeling, making it challenging to adapt effectively to diverse inputs. Sun et al. (2024) proposed the Diverse Size Experts structure for each FFN layer, where each expert has a different parameter size to handle generating tasks of varying difficulty. However, they employ a uniform routing strategy that fails to route input tokens to the most suitable expert, resulting in inefficient parameter utilization and compromised perfor-

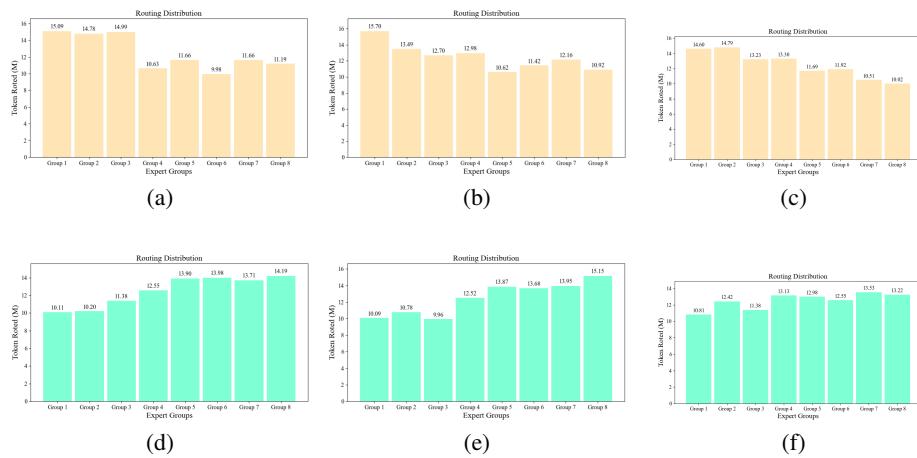


Figure 3: The number of tokens routed to each expert group. (a) Our MoHGE-1B with Group-Wise Auxiliary Loss. (b) Our MoHGE-3B with Group-Wise Auxiliary Loss. (c) Our MoHGE-14B with Group-Wise Auxiliary Loss. (d) Our MoHGE-1B without Group-Wise Auxiliary Loss. (e) Our MoHGE-3B without Group-Wise Auxiliary Loss. (f) Our MoHGE-14B without Group-Wise Auxiliary Loss.

455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1598  
1599  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1698  
1699  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1788  
1789  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1798  
1799  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1888  
1889  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1898  
1899  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1988  
1989  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1998  
1999  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2088  
2089  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2098  
2099  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138

486 REFERENCES  
487

488 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
489 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
490 report. *arXiv preprint arXiv:2303.08774*, 2023.

491 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
492 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

493 Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical com-  
494 monsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*,  
495 volume 34, pp. 7432–7439, 2020.

496 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,  
497 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are  
498 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

499 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
500 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to  
501 solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

502 OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models,  
503 2023.

504 William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter  
505 models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39,  
506 2022.

507 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and  
508 Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint  
509 arXiv:2009.03300*, 2020.

510 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,  
511 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL  
512 <https://arxiv.org/abs/2103.03874>, 2024.

513 Quzhe Huang, Zhenwei An, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Kun Xu,  
514 Liwei Chen, Songfang Huang, and Yansong Feng. Harder task needs more experts: Dynamic  
515 routing in MoE models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings  
516 of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long  
517 Papers)*, pp. 12883–12895, Bangkok, Thailand, August 2024. Association for Computational Lin-  
518 guistics. doi: 10.18653/v1/2024.acl-long.696. URL <https://aclanthology.org/2024.acl-long.696/>.

519 Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of  
520 local experts. *Neural computation*, 3(1):79–87, 1991.

521 Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm.  
522 *Neural computation*, 6(2):181–214, 1994.

523 Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly  
524 supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

525 Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child,  
526 Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language  
527 models. *arXiv preprint arXiv:2001.08361*, 2020.

528 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large  
529 language models are zero-shot reasoners. *Advances in neural information processing systems*,  
530 35:22199–22213, 2022.

531 Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong  
532 Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-  
533 of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

540 Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong  
 541 Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-  
 542 of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024b.

543

544 Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi,  
 545 Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset:  
 546 Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

547

548 Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Common-  
 549 sense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

550

551 Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton,  
 552 and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.  
 553 *arXiv preprint arXiv:1701.06538*, 2017.

554

555 Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan  
 556 Catanzaro. Megatron-Im: Training multi-billion parameter language models using model par-  
 557 allelism. *arXiv preprint arXiv:1909.08053*, 2019.

558

559 Manxi Sun, Wei Liu, Jian Luan, Pengzhi Gao, and Bin Wang. Mixture of diverse size experts. In  
 560 Franck Dernoncourt, Daniel Preoțiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the  
 561 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp.  
 562 1608–1621, Miami, Florida, US, November 2024. Association for Computational Linguistics.  
 563 doi: 10.18653/v1/2024.emnlp-industry.118. URL <https://aclanthology.org/2024.emnlp-industry.118/>.

564

565 Neil C Thompson, Kristjan Greenewald, Keeheon Lee, Gabriel F Manso, et al. The computational  
 566 limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10, 2020.

567

568 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée  
 569 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and  
 570 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

571

572 An Wang, Xingwu Sun, Ruobing Xie, Shuaipeng Li, Jiaqi Zhu, Zhen Yang, Pinxue Zhao, JN Han,  
 573 Zhanhui Kang, Di Wang, et al. Hmoe: Heterogeneous mixture of experts for language modeling.  
 574 *arXiv preprint arXiv:2408.10681*, 2024.

575

576 Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yo-  
 577 gatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language  
 578 models. *arXiv preprint arXiv:2206.07682*, 2022.

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594 

## 7 APPENDIX

595 

### 7.1 MODEL CONFIG

596 Detailed architectural configurations for all evaluated models are summarized in Table 5.  
597

600 Configuration	1B Scale	3B Scale	14B Scale
<b>Shared Configuration</b>			
601 Transformer Layers	9	15	36
602 Input Dim	1024	1024	1024
603 Attention Heads	16	16	16
<b>Dense Model</b>			
604 FFN Hidden Dim	4096	6144	8192
<b>MoE Baseline</b>			
606 $N_e$	32	64	128
607 $K_e$	6	6	6
608 Shared Experts $N_s$	2	2	2
609 Expert Hidden Dim	832	1024	1280
<b>MoHGE</b>			
610 $N_g$	8	8	8
611 $K_g$	3	3	3
612 $N_e$	32	64	128
613 $K_e$	6	6	6
614 Shared Experts $N_s$	2	2	2
615 Hidden Dims of Expert Groups	{256, 320, 384, 512, 640, 768, 832, 896}	{384, 512, 640, 768, 896, 1024, 1152, 1280}	{640, 768, 896, 1024, 1152, 1280, 1408, 1536}

615 Table 5: Architecture configurations of the evaluated models at both 1B, 3B and 14B parameter  
616 scales. MoHGE uses heterogeneous expert groups with different hidden dimensions.  
617618 

### 7.2 ROUTING ANALYSIS OF TOKENS OF DIFFERENT DIFFICULTIES

621 Token Ranks	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
622 Top 1K	16.3%	14.9%	14.1%	12.2%	12.3%	10.5%	10.0%	9.7%
623 Top 1K-5K	15.0%	14.4%	13.4%	13.1%	12.4%	10.3%	10.9%	10.5%
624 Top 5K-10K	13.6%	13.4%	13.7%	12.6%	11.5%	12.4%	11.5%	11.3%
625 Beyond 10K	11.3%	12.0%	11.4%	12.7%	13.0%	12.7%	13.6%	13.3%

626 We categorized the vocabulary into four difficulty levels based on occurrence frequency ranks in  
627 training corpus: Top 1K (easiest), Top 1K-5K, Top 5K-10K and Beyond 10K (most difficult). Sec-  
628 tion 7.2 shows the ratios of tokens with different difficulty routed to different expert groups. These  
629 results demonstrate that simpler tokens tend to be routed to expert groups with fewer parameters,  
630 and this validates the effectiveness of our method.  
631