Towards a Comprehensive Scaling Law of Mixture-of-Experts

Anonymous ACL submission

Abstract

Mixture-of-Experts (MoE) has emerged as the dominant paradigm for parameter-efficient scaling and cost-effective deployment of Large Language Models (LLMs). However, scaling laws developed for dense models are fundamentally inapplicable to MoE architectures, owing to three critical challenges: the multiplicity of governing factors, their intricate interdependencies, and the non-monotonic nature of their effects on model performance. These challenges collectively demand a rigorous and fine-grained investigation into MoE-specific scaling laws. In this work, we present a systematic decomposition of MoE configurations, identifying five key factors that influence model performance from both size and structural perspectives: data size (D), total model size (N), activated model size (N_a), number of active experts (G), and ratio of shared experts (S). We conduct 450 controlled experiments to characterize the marginal effects of each factor, culminating in a comprehensive and precise joint MoE scaling law. Building on this foundation, we derive the theoretically optimal and practically efficiency-aware optimal configurations for G, S and N_a/N with detailed analyses. Our findings reveal that the optimal settings for G and S are invariant to model and data size, while the optimal activation ratio N_a/N grows sparser as N increases. Our proposed MoE scaling law provides accurate and insightful guidance for future MoE model design and training.

1 Introduction

Large Language Models (LLMs) have been broadly adopted across diverse domains, demonstrating strong capabilities language understanding, reasoning, and agentic tasks. A particularly striking observation is that LLMs can continually extend their capability boundaries as model size and training data scale up. Scaling laws for LLMs (Kaplan et al., 2020; Hoffmann et al., 2022; Sun et al., 2025; Tian et al., 2025), which predict model loss as a function of key factors such as data and model size prior to training, provide principled guidance for selecting appropriate model configurations and training settings, and for maximizing model capability under given computational budgets or resource constraints. Recently, **Mixture-of-Experts (MoE)** has emerged as one of the mainstream architectures widely adopted in powerful industry-level LLMs (Liu et al., 2024, 2025; Qwen Team et al., 2025; OpenAI et al., 2025; Zeng et al., 2026; DeepSeek-AI, 2026). Different from dense architectures, which involve all parameters in the forward pass, employs multiple expert modules alongside a router that dynamically selects which experts to activate for each input token. This sparse activation mechanism allows MoE models to substantially benefit from increased total model size while preserving efficient inference.

Driven by the pursuit of more efficient LLMs, considerable efforts have been devoted to advancing MoE architectures. Shared experts have been introduced to robustly capture general knowledge (Dai et al., 2024; Sun et al., 2024), and a clear trend toward larger numbers of both activated and total experts has emerged in recent models (DeepSeek-AI, 2026; Team et al., 2026). However, existing scaling laws, whether developed for dense models (Kaplan et al., 2020; Hoffmann et al., 2022) or prior MoE models (Krajewski et al., 2024; Wang et al., 2024b; Abnar et al., 2025; Li et al., 2025a), fail to accurately predict performance under these evolving MoE configurations. This gap underscores the urgent need for a new MoE scaling law capable of reliably guiding model design and training.

To comprehensively characterize the key factors governing MoE model performance, we begin with the classical factors from dense scaling laws: **data size** (D) and **total model size** (N). Beyond these, the **activated model size** (N_a), which determines the parameters engaged during the forward pass, is an essential quantity unique to MoE models. From

the structural perspective, we identify the **number of activated experts** (G) as another critical factor, since N_a and G jointly determine the per-expert size. Furthermore, the **ratio of shared experts among activated experts** (S) is explicitly incorporated, as S and G together specify the exact counts of shared and routed experts. Our MoE scaling law is built upon these five factors. Constructing such a scaling law faces three key challenges: (a) Our MoE scaling law considers more comprehensive factors D, N, N_a, G, S compared to existing scaling laws. (b) Preliminary experiments imply that certain factors exhibit non-monotonic effects on loss, making them considerably harder to model. (c) These factors are mutually coupled, which significantly compounds the difficulty of deriving a unified joint scaling law.

To accomplish our MoE scaling law, we systematically identify five essential factors (D, N, N_a, G, S) for MoE models, and conduct 450 controlled experiments to record the corresponding validation losses $\mathcal{L}$. Based on these results, we first establish the base scaling law formulation with the fundamental factors D and N following (Hoffmann et al., 2022). Building upon this, we characterize the marginal effects of N_a, G, and S individually. By progressively investigating the relationships between each newly introduced factor and the hyperparameters of the joint scaling law, we ultimately arrive at the joint MoE scaling law, as follows:

$$L(N, D, N_a, G, S) = \Phi(G, S) \cdot \Psi(N, N_a) + \frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon \quad (1)$$

where $\Phi(G, S) = eG + \frac{f}{G} + mS^2 + nS$ is the structural term and $\Psi(N, N_a) = \frac{1}{N^\alpha} + \frac{k}{N_a^\alpha} + h\frac{N_a}{N}$ is the scale term. Our scaling law could satisfactorily predict MoE models' losses with larger data/model sizes (e.g., up to 30B N, 6.6B N_a and 100B D) and different MoE settings (e.g., up to 40 activated experts G and 70% ratio of shared experts S). Therefore, we contribute a scientifically rigorous step-by-step methodology for scaling law research, with high transparency and reproducibility, fostering the growth of the MoE scaling law community.

Based on our MoE scaling law, we conduct in-depth analyses and derive the following implications: (a) ***The optimal number of activated experts is around 7*** for typical MoE configurations. (b) ***The optimal activated parameter ratio*** N_a/N ***becomes increasingly sparse as*** N ***scales up.*** The theoretically optimal N_a/N ranges from 33.16% to 18.63% as N increases from 80B to 1.6T. Considering the cost, the practical efficiency-aware optimal ratios even narrow to 5%. (c) ***The presence of shared experts is essential, with their impact becoming increasingly pronounced as*** N ***scales up.*** We believe that our proposed MoE scaling law provides a more comprehensive understanding and more accurate prediction across diverse MoE configurations, and hope to facilitate LLM community in future MoE model design and training.

2 Preliminary

2.1 MoE Architecture

The MoE architecture replaces the dense FFN layers in standard Transformers with multiple independent experts, where only a small subset is activated per token (Fedus et al., 2022; Zhou et al., 2022; Jiang et al., 2024). This design enables model scaling by increasing the number of experts while maintaining nearly constant computational cost, making MoE an efficient and widely adopted framework for training and deploying large-scale LLMs.

2.2 Existing Scaling Laws of LLMs

Scaling Laws of Dense LLM. Scaling laws describe the relationship between key factors such as total model size N, data size D and the loss L. Classical scaling laws include the Chinchilla scaling law (Hoffmann et al., 2022), which states that L follows a power-law dependence on N and D, written as $L(N, D) = a/N^\alpha + b/D^\beta + \epsilon$. It consists of three terms: the first and second terms capture the limitations imposed by finite model size and finite data size, respectively. The last term, ϵ, represents the irreducible error that arises from the inherent uncertainty in the training data.

Scaling Laws for MoE. Unlike dense models, MoE introduces new structures with additional factors (e.g., N_a, G, S, etc), whose effects on model loss are non-monotonic and often interdependent. Existing scaling laws for dense models are insufficient to guide MoE's model design, which motivates the development of scaling laws tailored to MoE with these new-added factors. Recent studies have investigated MoE scaling laws based on certain MoE-specific factors, including the granularity of activated experts (Krajewski et al., 2024) and the activated model size (Ludziejewski et al., 2025). Different from them, our scaling law is more comprehensive and quantitatively defined with in-depth

correlations and insights among five factors. As a result, our scaling law provides a more accurate fit to the loss, as shown in Figure 5.

3 Experimental Setup

Key Factors. We systematically investigate five key factors governing MoE training: D, N, N_a, G, S. To isolate their individual effects, we conduct 450 controlled experiments organized into several groups, where each group varies a single target factor while keeping all others fixed. This controlled setup enables a rigorous assessment of each factor's marginal effect on validation loss.

Formally, we define the number of shared experts as n_s, the number of routed experts as n_e, the number of activated routed experts as n_k, the head dimension as d_{head}, the hidden dimension as d_{hidden}, the expert dimension as d_{expert}, the number of heads as n_h and the number of layers as l. Based on these definitions, the following relationships hold:

$$G = n_k + n_s, \quad S = n_s/G \tag{2}$$

$$N_a \approx \left(4d_{\text{head}} \cdot n_h + 3Gd_{\text{expert}}\right) \cdot d_{\text{hidden}} \cdot l \tag{3}$$

$$N \approx \left(4d_{\text{head}} \cdot n_h + 3d_{\text{expert}}(SG + n_e)\right) \cdot d_{\text{hidden}} \cdot l \tag{4}$$

Model and Training Configuration. An MoE layer consists of multiple experts and a router that assigns tokens, often using a Top-K routing strategy with an auxiliary balance loss to ensure expert utilization. Training typically adopts standard optimizers such as AdamW (Kingma and Ba, 2014) with parallelism techniques (data, model and expert parallelism) for scalability. We select the widely-used classical MoE structures with architectural details provided in Appendix B. All experiments employ the Warmup-Stable-Decay (WSD) learning rate scheduler (Hu et al., 2024). For studies involving different values of D, we reused the same warmup and stable phases across runs to avoid redundant computation and reduce resource usage. All models are trained on a subset of the Dolma V1.7 dataset (Soldaini et al., 2024).

Fitting Details. To fit the correlations between validation loss and five key factors, we systematically conduct experiments within controlled ranges, with total model size $N \in [133\text{M}, 3.4\text{B}]$ and training data sizes $D \in [10\text{B}, 50\text{B}]$ tokens. We additionally vary the activated model size $N_a \in [30\text{M}, 2.2\text{B}]$, the number of activated experts $G \in (1, 20]$ and the ratio of shared experts $S \in [0.0, 0.8]$. For validation, we further extend our experiments to larger model and data scales as well as different MoE configurations, thereby verifying the effectiveness and generalization capability of the proposed scaling laws. Specifically, the validation data points cover both interpolation and extrapolation, including $G \in (1, 40]$, $S \in [0.0, 0.7]$, $N_a \in [453\text{M}, 6.6\text{B}]$, $N \in [2.4\text{B}, 30\text{B}]$, and $D \in [10\text{B}, 100\text{B}]$. The complete experimental configurations are provided in Appendix K.

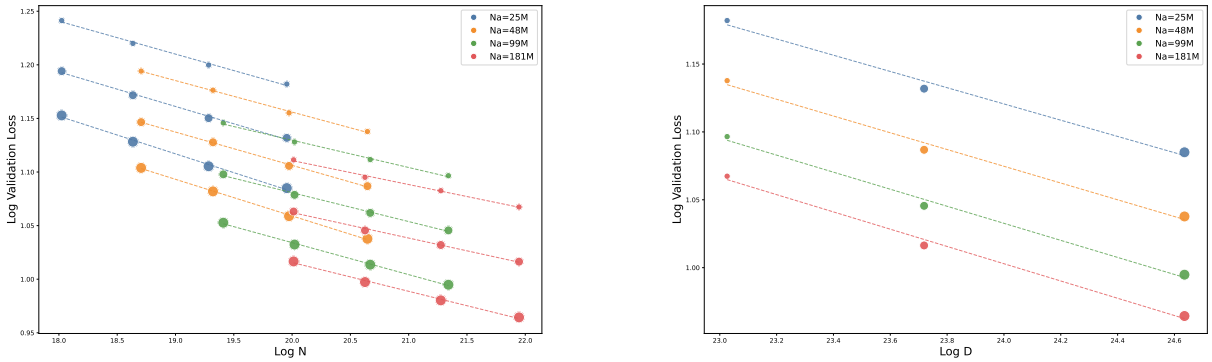
Specifically, we extract data from the decay phase of training logs, recording $(D, \mathcal{L})$ pairs, and apply a sliding window of size 10 for smoothing before selecting the validation loss at the final step. All factors are normalized to dimensionless values. We then predefine the scaling law formulation and perform least-squares fitting via `scipy.optimize.curve_fit`. Initial values and bounds are set following prior work and empirical heuristics. We additionally conduct a grid search to ensure that fitted parameters do not converge to boundary values, and observe consistent convergence across diverse initializations. Further details are provided in Appendix C.

Therefore, we contribute a scientifically rigorous, step-by-step methodology for scaling law research, characterized by high transparency and reproducibility. Our methodology encompasses extensive experiments and data collection, marginal-effect analysis for functional form characterization, hyperparameter independence analysis for deriving the joint scaling law, final function fitting with empirical verification, and the systematic derivation of core insights. We hope our work serves as a solid foundation for the broader MoE scaling law research community.

4 Our Scaling Laws for MoE

MoE has gradually emerged as a primary solution for the continuous scaling of model sizes and efficient deployment of LLMs. The core factors of MoE models exhibit higher complexity and strong inter-factor coupling compared to dense models and such inherent complexity renders the classical Chinchilla (Hoffmann et al., 2022) and OpenAI (Kaplan et al., 2020) scaling laws insufficient to guide the design of MoE architectures.

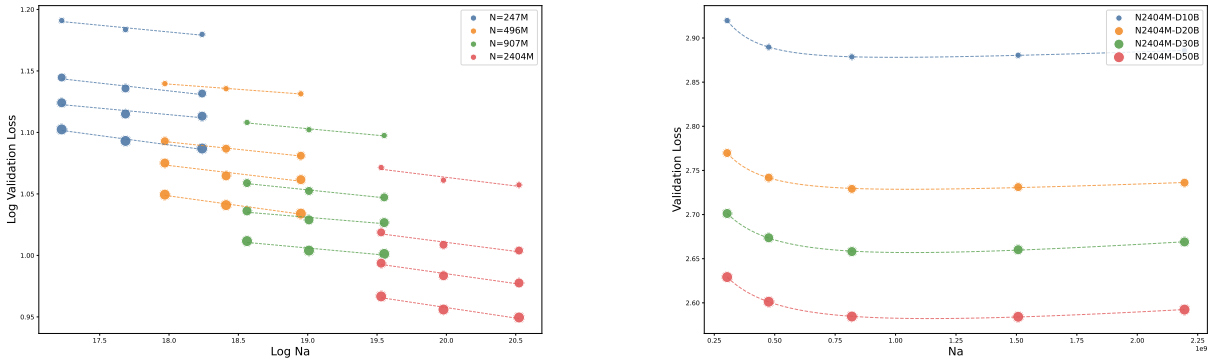
We aim to build a comprehensive and accurate MoE scaling law through a systematic pipeline: extensive experiments and data collection, marginal-effect analysis for functional form characterization,



(a) Loss vs. N.

(b) Loss vs. D.

Figure 1: Marginal effects of validation loss with respect to N and D under the logarithmic coordinate system. Data point sizes are directly proportional to D.



(a) Loss vs. N_a at small N_a/N ratios.

(b) Loss vs. N_a across the entire N_a/N.

Figure 2: Marginal effects of validation loss with N_a. (a) illustrates the power-law-like marginal relationship between loss and N_a under smaller N_a/N. (b) indicates that loss oscillates and increases as N_a/N becomes increasingly large. Data point sizes are proportional to D.

hyperparameter independence analysis for deriving the joint scaling law, and final function fitting with empirical verification. Throughout this process, we follow Occam's Razor, favoring simpler formulations with fewer hyperparameters while maintaining strong fitting performance, ensuring the validity and transferability of the resulting scaling law. In the following, we sequentially elaborate on each core factor on MoE model performance via controlled-factor experiments, encompassing N, D, N_a, G, and S.

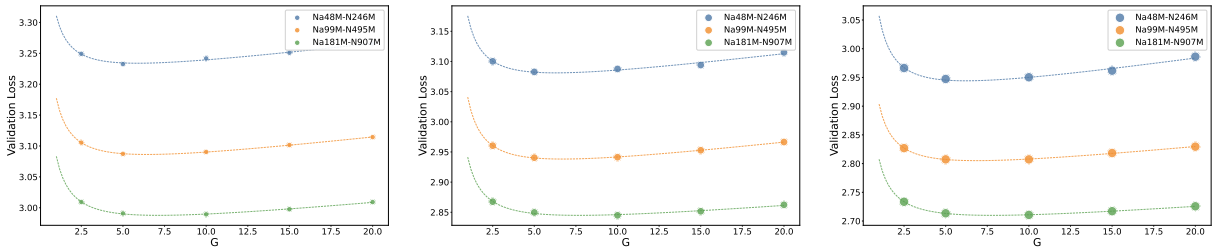
4.1 The Basic MoE Scaling law's form with N and D

Total model size N and data size D constitute two primary factors influencing the performance of LLMs. Leveraging the scaling laws of dense models as a foundation, we examine whether N and D in MoE still conform to a power-law relationship. As illustrated in Figure 1, a distinct power-law relationship is observed between validation loss, total model size N and data size D.

Specifically, we increase N with G, S, D and N_a unchanged. By performing logarithmic transformations on both the validation loss and N, a linear relationship is observed across different ranges of model sizes—this confirms a significant power-law relationship between the loss and N. Similarly, a significant power-law relationship also exists between the loss and D. Furthermore, in Figure 1, we find that experimental groups with the same model size but varying data sizes exhibit a translation along the y-axis, which implies that N and D are mutually independent. From these observations, we conclude that the loss for MoE with respect to the total model size N and data size D is as:

$$L(N, D) = \frac{a}{N^{\alpha}} + \frac{b}{D^{\beta}} + \epsilon \tag{5}$$

which has the same form of the Chinchilla scaling law (Hoffmann et al., 2022). The specific fitting results of Eq. 5 are provided in Appendix C.2.



(a) Loss vs. G with 10B data size. (b) Loss vs. G with 20B data size. (c) Loss vs. G with 50B data size.

Figure 3: Marginal effects of validation loss with respect to G. (a), (b) and (c) illustrate the marginal relationship between loss and G under different D and N. Data point sizes are proportional to D. Note that non-integer values of G (e.g., $G = 2.5$) denote corresponds to 2 activated routed experts and 0.5 shared experts.

4.2 Impact of Activated Model Size N_a

Activated model size N_a is a critical factor specific to MoE architectures that governs the balance between the model performance and efficiency. To gain deeper insights into the scaling law with N_a as an independent factor, we conducted multiple controlled experiments where N_a was the only varying factor. The formula followed by the controlled variable of N_a is detailed in Appendix H.

In Figure 2, the scaling of activated parameters N_a is achieved by increasing the expert dimension, while the total model size N is maintained constant through a corresponding reduction in the number of routed experts. When the ratio of N_a to N is small, a power-law-like relationship is exhibited between N_a and validation loss. However, as this ratio increases, the validation loss demonstrates a tendency to rise gradually. An overall distribution resembles a hook-like function, which is formalized as follows:

$$L(N_a) = \frac{c}{N_a^{\gamma}} + hN_a + \iota \tag{6}$$

Next, we proceed to integrate the relationship involving N_a with those of data size D and total model size N. We performed hyperparameter fitting under different configurations of D and N, with the results presented in Figure 9. We observe that ι exhibits a negative correlation with both D and N, following a power-law distribution. In contrast, c and γ exhibit oscillations with variations in N and D, indicating that they bear no systematic relationship to N and D. Furthermore, h is negatively correlated with N, exhibiting an inversely proportional relationship, while showing no dependence on D. Therefore, the joint scaling law of N, D and N_a is concluded as follows:

$$L(N, D, N_a) = \frac{c}{N_a^{\gamma}} + h\frac{N_a}{N} + \iota L(N, D) \tag{7}$$

$L(N, D)$ denotes the basic scaling law sharing the same form in Eq. 5. Notably, our fitting results reveal that $\gamma \approx \alpha$. Hence, the final scaling law that governs $L(N, D, N_a)$ is formalized as follows:

$$L(N, D, N_a) = \frac{a}{N^{\alpha}} + \frac{b}{D^{\beta}} + \frac{c}{N_a^{\alpha}} + h\frac{N_a}{N} + \epsilon \tag{8}$$

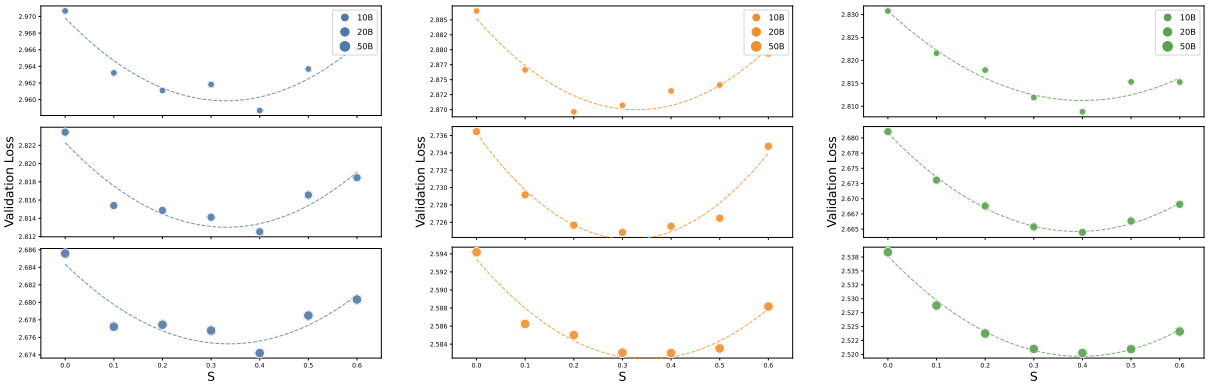
We also explored alternative relationship forms of $L(N_a)$, performed relevant comparative experiments and derivations and ultimately established the aforementioned formulas describing the scaling law of N, D and N_a.

Eq. 8 implies the existence of an optimal value for N_a. Intuitively, when N_a/N exceeds a certain threshold while other factors remain fixed, the per-expert size increases progressively as the number of experts decreases. This leads to a gradual structural distortion in the MoE architecture, which undermines the combinatorial activation advantage of MoE and consequently degrades model performance. We further validate that Eq. 8 generalizes accurately to models with larger sizes. Further discussions and detailed fitting procedures are provided in Appendices C.3 and F.

4.3 Impact of the Number of Activated Experts G

G constitutes another critical factor, defined as the number of activated experts (including activated shared and routed experts). It reflects the granularity of expert partitioning in MoE. When all other factors are held fixed, scaling G is equivalent to adjusting the granularity of expert partitioning.

To investigate the scaling law with G as the independent factor, a series of controlled experiments were conducted with other factors constant, shown in Figure 3. With the increase in G, the validation loss exhibits a trend of first decreasing and then increasing. It is hypothesized that the impact of



(a) Loss vs. S: Na240M-N1.2B. (b) Loss vs. S: Na476M-N2.4B. (c) Loss vs. S: Na793M-N4B.

Figure 4: Marginal effects of validation loss with respect to S. (a), (b) and (c) respectively illustrate the marginal relations between loss and S under different N, D settings. For small N and D, the trend of S is taking shape but is also easily affected by noise. As they increase, the scaling law of S gradually becomes significant and robust. More details are in Appendix C.5.

G better conforms to a hook function relationship, expressed as follows:

$$L(G) = eG + \frac{f}{G} + \tau \tag{9}$$

It is noteworthy that the fitted exponent of G approaches 1. Therefore, following the Occam's Razor principle, the exponent term of G is omitted. Other functional forms are also attempted but yielded inferior fits.

4.4 The Joint Scaling Law of N, D, N_a, G

Building on Eq. 8 and Eq. 9, we examine how the fitted hyperparameters $a, \alpha, b, \beta, c, h, \epsilon$ vary across different values of G. We observe that a, c, and h follow a hook-function trend with respect to G, while the remaining hyperparameters are largely insensitive to G and exhibit no discernible pattern, as shown in Figure 10. Accordingly, we parameterize them as $a = e_1 G + \frac{f_1}{G} + \tau_1$, $c = e_2 G + \frac{f_2}{G} + \tau_2$, and $h = e_3 G + \frac{f_3}{G} + \tau_3$. Upon re-parameterization, we further find that the pairs (e_1, f_1), (e_2, f_2), and (e_3, f_3) exhibit proportional relationships, and that $\tau_3 \approx 0$. The resulting scaling law for $L(N, D, N_a, G)$ is thus as follows:

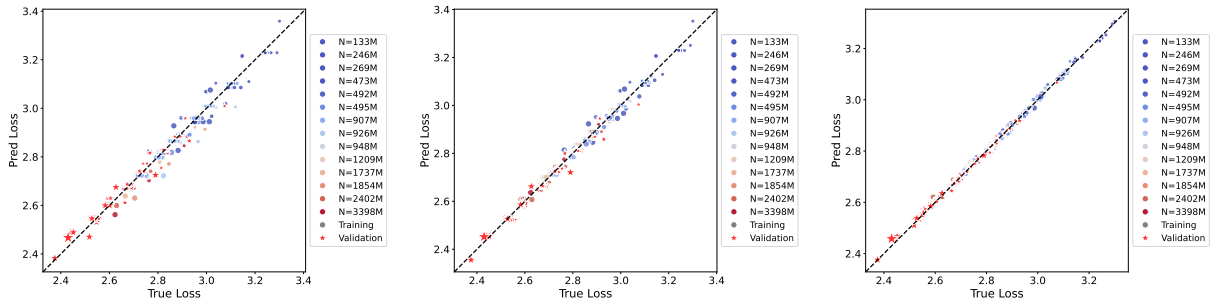
$$\begin{aligned} L(N, D, N_a, G) = (eG + \frac{f}{G}) \cdot \Psi(N, N_a) \\ + \frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon \end{aligned} \tag{10}$$

Here, $\frac{a}{N^\alpha} + \frac{b}{D^\beta} + \epsilon$ corresponds to the Chinchilla-like base scaling law from Eq. 5. Since N_a and N exert similar effects on the loss to a certain extent, the term $\frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon$ can be interpreted as characterizing the scaling behavior with respect to both (activated and total) model size and data size. The term $eG + \frac{f}{G}$ captures the effect of G on loss (reflecting the MoE expert structure), and this effect is modulated by the scale term $\Psi(N, N_a) = \frac{1}{N^\alpha} + \frac{k}{N_a^\alpha} + h\frac{N_a}{N}$, which exhibits non-monotonic behavior with respect to N_a as introduced in Eq. 8. Detailed analyses of the hyperparameters and fitting results for Eq. 10 are provided in Appendix C.4.

4.5 Extended Joint MoE Scaling Law with Shared Expert Ratio S

Shared experts have been shown to be essential in popular MoE architectures (Dai et al., 2024; Zeng et al., 2026). We vary the shared expert ratio S ($S = n_s/G$, where n_s denotes the number of shared experts) as the sole varying factor and conduct the corresponding experiments, with results shown in Figure 4. The results confirm that MoE models with shared experts significantly outperform those without, validating the necessity of shared expert isolation. Furthermore, the influence of S becomes increasingly pronounced as N scales up. As S increases, the loss first decreases and then increases, exhibiting a clear non-monotonic trend that closely follows a quadratic pattern. Moreover, the fitted curves across different settings are approximately symmetric about the same axis of symmetry. Based on these observations, we adopt a quadratic function to characterize the marginal effect of S on the loss.

$$L(S) = mS^2 + nS + \psi \tag{11}$$



(a) (Krajewski et al., 2024). (b) (Abnar et al., 2025). (c) Ours.

Figure 5: Fitting results of existing and our joint scaling laws for MoE architectures. The average validation loss errors of (a), (b) and (c) are respectively 0.0193, 0.0170 and 0.0060. Star points are validation data with larger model sizes, larger data sizes and different MoE settings. Data point size is proportional to D. Detailed descriptions of compared scaling laws are provided in Appendix I.

Final Joint MoE Scaling Law. Following the same approach as G, we incorporate S as the second structural factor into Eq. 10 to construct the final joint MoE scaling law over all five factors. To this end, we fit the hyperparameters of Eq. 11 across diverse configurations of N, D, N_a, and G, and analyze their relationships with each factor. The results are presented in Figure 12, yielding the following key findings: (1) m and n are independent of D, indicating that S and D are mutually decoupled; (2) m increases while n decreases as N and N_a grow; notably, the extremum of S remains invariant to changes in N and N_a; (3) ψ exhibits a clear power-law relationship with N, N_a, and D; (4) since S is already correlated with G by definition (Eq. 4), its dependence on G is not further modeled. To ensure compatibility with Eq. 10, the final form of $L(N, D, N_a, G, S)$ is expressed as $\Phi(G, S) \cdot \Psi(N, N_a) + \frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon$. The complete joint MoE scaling law over all five factors N, D, N_a, G, S is thus presented as follows:

$$L(N, D, N_a, G, S) = \Phi(G, S) \cdot \Psi(N, N_a) + \frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon \tag{12}$$

Note that the structural factors S and G, which jointly characterize the activated expert configuration of the MoE architecture, are subsumed in $\Phi(G, S)$. Specifically, $\frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon$ captures the scaling behavior of model loss with respect to model and data sizes. The structural term $\Phi(G, S)$ and the size-related term $\Psi(N, N_a)$ are multiplicatively coupled, jointly governing the overall MoE scaling law. Detailed hyperparameter analyses are provided in Appendix C.5.

Fitting Results. To determine the hyperparameters in Eq. 12, we utilize all 450 experiments spanning diverse configurations of N, D, N_a, G, and S. The fitted hyperparameter values are reported in Table 2 in Appendix C.1. We then evaluate the fitting performance of our MoE scaling law in Figure 5. To ensure a fair comparison, all baseline scaling laws are re-fitted using the same set of data points before prediction, and our experimental data fully subsumes those used by the baselines. The results demonstrate that our scaling law achieves superior fitting performance over existing alternatives. Furthermore, we extend the evaluation to larger MoE models (up to 30B N, 6.6B N_a and 100B D) across a broad range of structural settings ($G \in (1, 40]$ and $S \in [0, 0.7]$). The consistently accurate predictions confirm that our scaling law generalizes robustly to larger-scale MoE models across wide ranges of configurations.

5 Key Implications for MoE Design

In this section, we discuss key insights derived from our MoE scaling law, offering practical guidance for effective MoE model design.

5.1 Implication-1: Optimal Number of Activated Experts G

In light of Eq. 12, the optimal number of activated experts G can be expressed as $G_{opt} = \sqrt{f/e}$, which shows that the optimal G is independent of model size N, activated model size N_a, and data size D, yielding a fixed optimal value of $G_{\text{opt}} \approx 6.78$. Notably, this theoretically derived optimum aligns closely with popular configurations adopted by mainstream MoE, including DeepSeek-V4 (DeepSeek-AI, 2026), Kimi-K2.6 (Team et al., 2026), and GLM5.1 (Zeng et al., 2026), indirectly implying the effectiveness of our implication.

5.2 Implication-2: Optimal Ratio of shared experts S

Similarly, S is independent of other factors and also has an optimal value as follows:

$$S_{opt} = -n/2m. \tag{13}$$

Eq. 13 shows that the optimal $S \approx 0.31$ is likewise independent of all other factors. The presence of shared experts yields substantial loss reduction. As S increases further, the loss continues to decrease marginally before rising sharply beyond the optimal point, with this effect becoming increasingly pronounced as N scales up. In summary, our findings lead to the following recommendation: shared experts are essential, and for the optimal activated expert number $G = 7$, setting 1 or 2 shared experts is advisable. This conclusion is almost consistent with DeepSeek (DeepSeek-AI, 2026; Liu et al., 2024). Further details are provided in Appendix E.

5.3 Implication-3: Optimal Activated Parameter Ratio N_a/N

Theoretical Analysis. Eq. 12 contains two types of terms involving the activated model size N_a: one where N_a appears in the denominator and one where it appears in the numerator. These two terms exert opposing effects on the model loss, implying the existence of an optimal N_a for any given configuration of the remaining factors (e.g., total model size N). Given N, the theoretically optimal activation ratio $\left(\frac{N_a}{N}\right)_{\text{opt}}$ is derived as follows (comprehensive derivation provided in Appendix F):

$$\begin{aligned}\left(\frac{N_a}{N}\right)_{opt_t} &= \left(\frac{\alpha \cdot [k \cdot \Phi(G,S) + c]}{hN^{\alpha} \cdot \Phi(G,S)}\right)^{\frac{1}{\alpha+1}} \\ &= \left(\frac{\alpha \cdot [k \cdot const + c]}{hN^{\alpha} \cdot const}\right)^{\frac{1}{\alpha+1}}.\end{aligned} \tag{14}$$

According to our Implications #1 and #2, G and S have optimal values and thus $\Phi(G,S) = eG + \frac{f}{G} + mS^2 + nS$ can be represented as a constant term $const$ under the optimal setting. Eq. 14 indicates that the optimal $(\frac{N_a}{N})_{opt_t}$ decreases as the model size N increases. It indicates that with the increasing total model sizes, the optimal MoE architecture will be sparser with smaller N_a, which is consistent with the current trend of MoE models (Team et al., 2026; OpenAI et al., 2025; DeepSeek-AI, 2026).

Practical Efficiency-aware Analysis. However, the theoretically optimal activation ratio $\frac{N_a}{N}$ derived in Eq. 14, cannot be directly applied to practical MoE architecture design, as computational efficiency is an equally critical consideration for real-world LLM deployment. Specifically, as N_a approaches its theoretical optimum, performance gains become increasingly marginal while the associated costs rise steadily, making it necessary to identify a practical, efficiency-aware optimal $\left(\frac{N_a}{N}\right)_{\text{opt}_e}$ that balances performance gain against efficiency cost.

Specifically, we define the loss gain threshold as ΔLoss for the step size of ΔN_a set as $0.01N$. As N_a is incrementally scaled for each step size, the marginal gain of loss reduction will ultimately fall below the loss gain threshold ΔLoss, where we suppose the model reaches the practical efficiency-aware optimal $(\frac{N_a}{N})_{opt_e}$. Comprehensive derivation and pseudo code are in Appendix F. Hence, for a given model size N, our MoE scaling law yields a practically applicable range for N_a, spanning the interval from the practical efficiency-aware optimal point to the theoretical optimal point, i.e., $N_a \in [(\frac{N_a}{N})_{opt_e}, (\frac{N_a}{N})_{opt_t}]$. To substantiate the validity of our conclusions, we conducted an analysis on the configurations of mainstream industrial MoE models, with detailed specifications in Table 4. It shows that the activated model sizes of most mainstream MoE models are within our recommended range above. Note that some recent MoE models (e.g., Kimi-K2.6 (Team et al., 2026)) employ a more aggressive sparser architecture with $\frac{N_a}{N} \leq 4\%$, primarily aiming to reduce the training/inference costs in practice with larger N. Nevertheless, our implication still helps to quantitatively measure the expected loss under different N_a, balancing effectiveness and efficiency.

6 Conclusion and Future Work

In this work, we propose a comprehensive joint MoE scaling law incorporating five key factors (N, D, N_a, G, and S), derived through systematic marginal effect analysis and hyperparameter independence analysis. Building on this, we further derive closed-form expressions for the loss-optimal configurations of the MoE-specific factors G, S, and N_a/N, and analyze their optima from both theoretical and practical perspectives, yielding several actionable insights. We believe that our proposed MoE scaling law with its insights could facilitate LLM community in future MoE model design and training. In the future, we will further validate our scaling law under larger scales and novel MoE architectures.

Limitations

Our work has several limitations. In this work, we mainly focus on the classical MoE architecture. The analysis has not been validated at extremely larger scales or with alternative MoE architectures / training objectives due to the resource limit. In addition, we focus primarily on MoE-related factors and do not examine other components of LLMs that may also impact the performance of MoE, such as the attention layers and their interactions. Since our work primarily focuses on MoE-related factors, we fix the batch size and learning rate to reasonable values. In future work, we plan to incorporate batch size and learning rate into a unified and comprehensive scaling law formulation.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin Mohamed Elnouby Ali, Josh Susskind, and Vimal Thilak. 2025. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models. *arXiv preprint arXiv:2501.12370*.

Mouxiang Chen, Binyuan Hui, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Jianling Sun, Junyang Lin, and Zhongxin Liu. 2025. Parallel scaling law for language models. *arXiv preprint arXiv:2505.10475*.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, and 1 others. 2022. Unified scaling laws for routed language models. In *International conference on machine learning*, pages 4057–4086. PMLR.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

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

Xu Owen He. 2024. Mixture of a million experts. *arXiv preprint arXiv:2407.04153*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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

Team Kimi, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, and 1 others. 2024. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Houyi Li, Ka Man Lo, Ziqi Wang, Zili Wang, Wenzhen Zheng, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. 2025a. Can mixture-of-experts surpass dense llms under strictly equal resources? *arXiv preprint arXiv:2506.12119*.

Houyi Li, Wenzhen Zheng, Jingcheng Hu, Qiufeng Wang, Hanshan Zhang, Zili Wang, Shijie Xuyang, Yuantao Fan, Shuigeng Zhou, Xiangyu Zhang, and 1 others. 2025b. Predictable scale: Part i–optimal hyperparameter scaling law in large language model pretraining. *arXiv e-prints*, pages arXiv–2503.

Houyi Li, Wenzhen Zheng, Qiufeng Wang, Zhenyu Ding, Haoying Wang, Zili Wang, Shijie Xuyang,

Ning Ding, Shuigeng Zhou, Xiangyu Zhang, and 1 others. 2025c. Farseer: A refined scaling law in large language models. *arXiv preprint arXiv:2506.10972*.

Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, and 1 others. 2024. Torchtitan: One-stop pytorch native solution for production ready llm pre-training. *arXiv preprint arXiv:2410.06511*.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, and 1 others. 2024. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ao Liu, Botong Zhou, Can Xu, Chayse Zhou, ChenChen Zhang, Chengcheng Xu, Chenhao Wang, Decheng Wu, Dengpeng Wu, and 1 others. 2025. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. *arXiv preprint arXiv:2505.15431*.

Jan Ludziejewski, Maciej Pióro, Jakub Krajewski, Maciej Stefaniak, Michał Krutul, Jan Małaśnicki, Marek Cygan, Piotr Sankowski, Kamil Adamczewski, Piotr Miłoś, and 1 others. 2025. Joint moe scaling laws: Mixture of experts can be memory efficient. *arXiv preprint arXiv:2502.05172*.

Ang Lv, Ruobing Xie, Yining Qian, Songhao Wu, Xingwu Sun, Zhanhui Kang, Di Wang, and Rui Yan. 2025. Autonomy-of-experts models. *arXiv preprint arXiv:2501.13074*.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. 2018. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*.

Sandhini OpenAI, Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R Fung, and 1 others. 2025. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*.

An Qwen Team, Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Xian Shuai, Yiding Wang, Yimeng Wu, Xin Jiang, and Xiaozhe Ren. 2024. Scaling law for language models training considering batch size. *arXiv preprint arXiv:2412.01505*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, and 1 others. 2024. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*.

Xingwu Sun, Shuaipeng Li, Ruobing Xie, Weidong Han, Kan Wu, Zhen Yang, Yixing Li, An Wang, Shuai Li, Jinbao Xue, and 1 others. 2025. Scaling laws for floating point quantization training. In *Proceeding of ICML*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, and 1 others. 2025. Longcat-flash technical report. *arXiv preprint arXiv:2509.01322*.

Changxin Tian, Kunlong Chen, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. 2025. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models. *arXiv preprint arXiv:2507.17702*.

An Wang, Xingwu Sun, Ruobing Xie, Shuaipeng Li, Jiaqi Zhu, Zhen Yang, Pinxue Zhao, JN Han, Zhanhui Kang, Di Wang, and 1 others. 2024a. Hmoe: Heterogeneous mixture of experts for language modeling. *arXiv preprint arXiv:2408.10681*.

Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. 2024b. Scaling laws across model architectures: A comparative analysis of dense and moe models in large language models. *arXiv preprint arXiv:2410.05661*.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. 2024. How does critical batch size scale in pre-training? *arXiv preprint arXiv:2410.21676*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Appendix

A The Use of Large Language Models

In this paper, we leveraged LLMs to support and refine the writing. Specifically, LLMs were used for grammar and spelling correction, as well as polishing linguistic expressions to enhance clarity and readability. All other core components of the work, including the development of ideas, design and execution of experiments and derivation of formulas, were completed manually by ourselves.

B Hyperparameter Details

We report serveral typical hyper-parameters used for training our MoE models in Table 1. The table includes only several typical MoE model configurations categorized by N. The models vary in layers, hidden size and expert size across different scales. All models are trained with AdamW and cosine learning rate decay, using a sequence length of 2048 and a batch size of 2M tokens. The detailed hyper-parameters of our MoE models are given as follows. In all of our experiments, the parameters considered do not include those from the embedding layer. For all experimental settings, refer to Appendix K.

C Fitting Details of Our MoE Scaling Laws

Our MoE scaling law precisely characterizes the influence of three critical dimensions on model performance: parameter scale, data size, and model architecture. Specifically, $\Phi(G, S) \cdot \Psi(N, N_a)$ captures the architectural contribution to loss and its modulation by parameter scale, while $\rho(N, D, N_a) = \frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon$ characterizes the joint effect of parameter and data scale on MoE loss. By unifying all core governing factors into a single coherent formulation, our scaling law provides a principled and theoretically grounded basis for MoE architecture design. Grounded in the true marginal and joint relationships among the five factors, we apply Occam's Razor to eliminate unnecessary hyperparameters wherever possible, ensuring both accuracy and the simplicity and transferability of the resulting joint scaling law. Details of the fitting procedure are presented below.

Fitting Details. We extract data from the decay phase of training logs as $(D,\ \mathcal{L})$ pairs, apply a sliding window of size 10 for smoothing, and select the validation loss at the final step. All factors and validation losses are normalized to dimensionless values. Based on the processed data, we predefine the scaling law formulations for both the baseline methods and our proposed law, and estimate the optimal parameters using `scipy.optimize.curve_fit`, a standard nonlinear least-squares fitting tool. Initial values and parameter bounds are set following prior work and empirical heuristics. To mitigate sensitivity to initialization, we conduct a grid search over the initial parameter space and enforce the criterion that fitted parameters must not converge to the predefined bounds. Across all reasonable initializations, the optimized hyperparameters consistently converge to similar values, confirming the robustness of the fitting procedure. To facilitate reproducibility, we release the fitting data, validation set (including N, N_a, D, G, S, and validation loss), and the corresponding Python scripts in the supplementary materials.

Importantly, given properly specified initial values and parameter bounds, the fitting procedure is highly robust: repeated experiments with different random seeds yield consistently stable results, indicating that additional uncertainty modeling for the fitted hyperparameters is unnecessary.

Furthermore, the comprehensive incorporation of relevant factors is a central objective of this work. Extensive validation across diverse factor configurations demonstrates that our joint scaling law consistently outperforms existing MoE scaling laws, underscoring the importance of accounting for a broad set of factors. Throughout the fitting process, we adhere to Occam's razor by favoring parsimonious functional forms over unnecessarily complex alternatives.

C.1 Numberical Fits of Our Joint MoE Scaling Law

Our joint MoE scaling law is formalized as:

$$L(N, D, N_a, G, S) = \Phi(G, S) \cdot \Psi(N, N_a) + \frac{a}{N^\alpha} + \frac{b}{D^\beta} + \frac{c}{N_a^\alpha} + \epsilon \tag{15}$$

where $\Phi(G, S) = eG + \frac{f}{G} + mS^2 + nS$ is the structural term and $\Psi(N, N_a) = \frac{1}{N^\alpha} + \frac{k}{N_a^\alpha} + h\frac{N_a}{N}$ is the scale term.

The fitted constants derived from 450 experiments across diverse model configurations are presented below. Of these, 268 experiments are used for fitting, 92 for validation, and 90 small-scale experiments for observing the marginal effect of

Table 1: Typical model hyper-parameters for different sizes. The table includes only several typical MoE model configurations categorized by N. The scaling law experiments corresponding to different factors are extended to these typical model configurations. Details of the parameter configurations can be found in Appendix K.

Total model size	247M	496M	907M	2.40B	3.96B
Activated model size	48M	99M	181M	476M	793M
# Layers	12	12	12	20	24
# Routed experts	32	32	32	32	32
# Activated routed experts	4	4	4	4	4
# Shared experts	1	1	1	1	1
# Attention heads	8	12	16	20	24
# Hidden size	512	768	1024	1280	1536
# Expert size	384	512	704	896	1024
# Attention head size	64	64	64	64	64
Optimizer	AdamW				
Adam (β_1, β_2)	(0.9, 0.95)				
Adam ϵ	1×10^{-8}				
Weight decay	0.1				
Clip grad norm	1.0				
Max lr	3.0×10^{-4}				
Min lr	0				
Lr decay	Cosine				
Decay rate	10%				
Sequence length	2048				
Batch size (# tokens)	2M				
Warmup steps	500				
Normalization	RMSNorm				
Vocabulary size	128256				
Positional encoding	ROPE				

G. For each factor, we systematically explore a wide range of candidate functional forms and retain those that best characterize its marginal effect on validation loss.

It is worth noting that, due to the presence of a constant term, none of the log-log curves presented below exhibit perfectly linear behavior. Nevertheless, an approximately linear trend in log-log space is sufficient to confirm a power-law relationship.

C.2 Fitting Results of the Scaling Law for $L(N, D)$

Figure 6 presents the fitting performance of $L(N, D)$ (Eq. 5), which considers only total model size N and data size D. As shown, $L(N, D)$ yields only a coarse-grained fit to our experimental data, with notably suboptimal accuracy on specific configurations. This demonstrates that additional MoE-specific factors must be taken into account to achieve precise performance prediction.

Table 2: Fitted constants and their values in Eq. 15.

Constant	Value
e	0.1577
f	7.2446
m	5.1395
n	-3.2363
k	0.0013
h	0.0450
a	38.0510
α	0.2383
b	27129.0488
β	0.4694
c	31.0958
ϵ	1.8182

C.3 Fitting Results of the Scaling Law for $L(N, D, N_a)$

Figure 7 presents the fitting performance of $L(N, D, N_a)$ (Eq. 8), which additionally incorporates the activated model size N_a. The results demonstrate that N_a is a critical factor for MoE

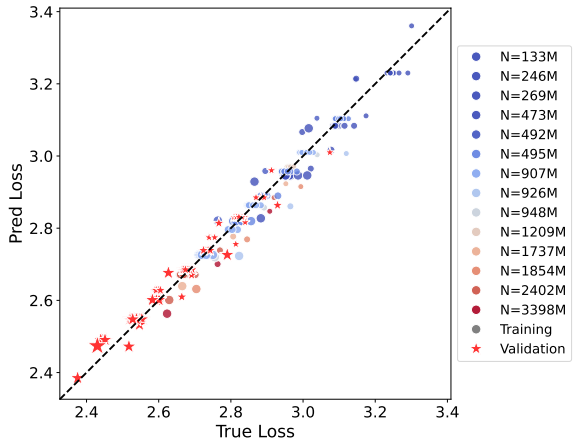


Figure 6: Fitting results of the scaling law of $L(N, D)$. Average validation loss error: 0.0195.

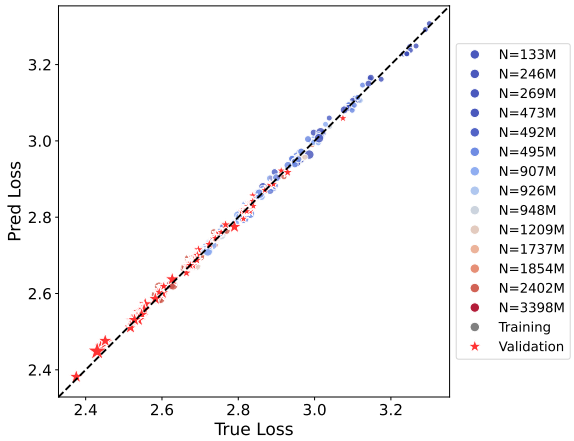


Figure 8: Fitting results of the scaling law of $L(N, D, N_a, G)$. Average validation loss error: 0.0085.

scaling, and its inclusion yields a substantial improvement in fitting accuracy over $L(N, D)$.

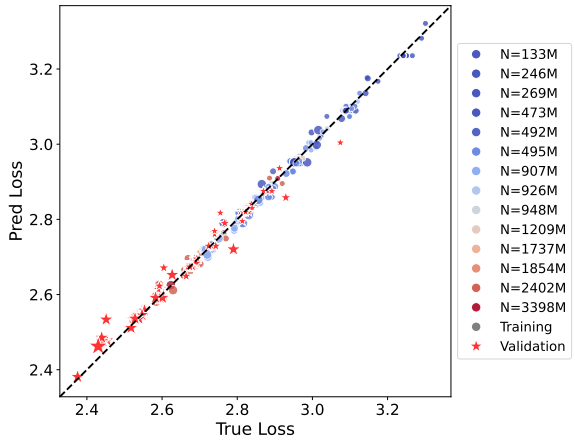


Figure 7: Fitting results of the scaling law of $L(N, D, N_a)$. Average validation loss error: 0.0143.

The process of incorporating N_a into the scaling law $L(N, D, N_a)$ (Eq. 8) is described as follows. We first design and conduct a series of controlled experiments on N_a following Eq. 36, and then perform hyperparameter fitting across diverse configurations of D and N. The associated results are presented in Figure 9.

It can be observed that ι exhibits a linear relationship with both N and D under logarithmic transformation, indicating a power-law dependence on each. The parameter h follows an inverse proportional relationship with N across different data volumes D. The remaining parameters c and γ fluctuate without displaying discernible correlations with N or D, and are therefore treated as independent of these factors. Notably, the fitted exponent of N_a in the term $h\frac{N_a}{N}$ approaches 1, providing justification for omitting the exponent term and adopting $h\frac{N_a}{N}$ in its current linear form.

C.4 Fitting Results of the Scaling Law for $L(N, D, N_a, G)$

Figure 8 presents the fitting performance of $L(N, D, N_a, G)$ (Eq. 10), which further incorporates the expert granularity G. As discussed in Section 4.3, G characterizes the structural properties of MoE architectures and captures their impact on model performance. The results demonstrate that incorporating the structural factor G yields a further improvement in fitting accuracy.

By analogy, to incorporate G into the scaling law $L(N, D, N_a, G)$, we first designed and conducted a series of controlled experiments where G was treated as the sole factor of variation, with all other factors held constant. Specifically, as G increases, the counts of routed experts and shared experts expand proportionally, whereas the corresponding expert dimensions shrink proportionally. Considering the marginal effect of G on loss and its coupling with N and N_a, we hypothesize that the hyperparameters a, c, and h, which appear as numerator, side coefficients—are correlated with G. This is further supported by the observation that D is independent of G. The variation of the fitted hyperparameters a, c, and h across different values of G is presented in Figure 10.

C.5 Fitting Results of the Scaling Law for $L(N, D, N_a, G, S)$

Similarly, we incorporate the shared expert ratio S (i.e., the ratio of shared to activated experts) as the final structural factor. The hyperparameter analysis

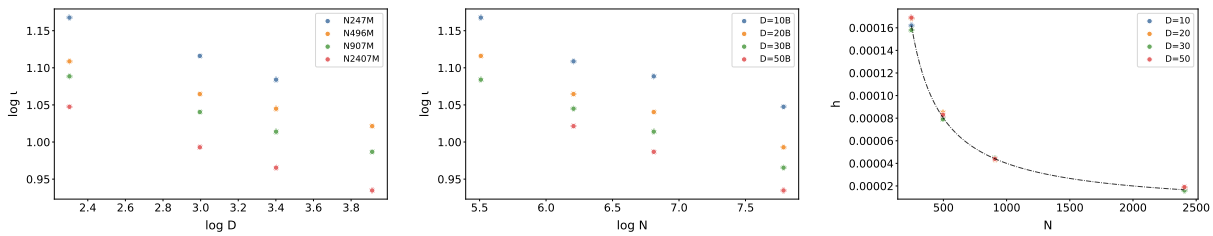


Figure 9: The correlations between ι, h in Eq. 6 and N, D. ι and h can be viewed as functions of N or D.

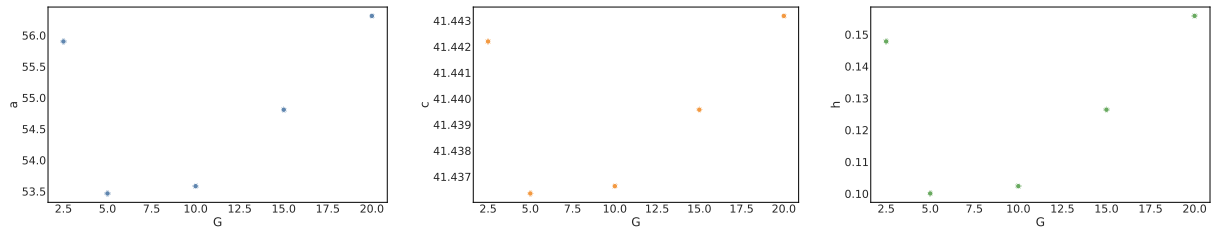


Figure 10: The correlations between a, c, h in Eq. 8 and G. a, c, h can be viewed as functions of G.

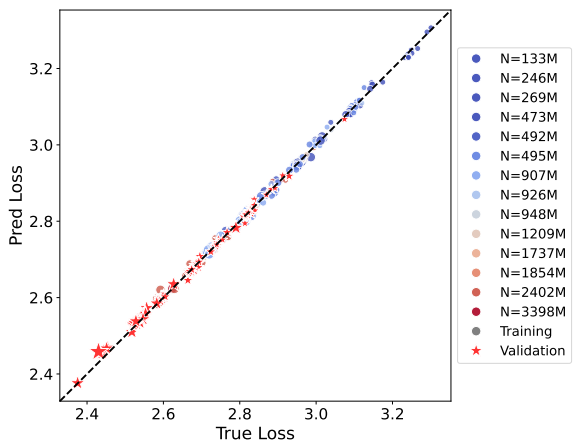


Figure 11: Fitting results of our final scaling law of $L(N, D, N_a, G, S)$. Average validation loss error: 0.0059. The fitting performance improves consistently as additional factors are incorporated.

with respect to S is presented in Figure 12, and the fitting performance of the complete joint scaling law $L(N, D, N_a, G, S)$ is illustrated in Figure 11. The analysis of hyperparameters related to S is presented in Figure 12. It can be observed: (1) m and n are independent of D, indicating that D and S are mutually decoupled; (2) m increases with the growth of N and N_a, whereas n decreases with the growth of N and N_a. Notably, the extreme point of S remains unchanged with variations in N and N_a; (3) ψ exhibits an obvious power-law relationship with N, N_a and D.

Furthermore, as stated in Section 4.5, the scaling law of S becomes increasingly prominent with the growth of model size. As shown in Figure 13.

D Derivation of Theoretical Optimal G

First, we decompose $L(N, D, N_a, G, S)$ Eq. 15 into three components to isolate terms involving G. This decomposition leverages the fact that most variables (N, D, N_a, S) are independent of G:

$$\begin{aligned} L_{N,D,N_a,S}(G) = & \underbrace{\left(eG + \frac{f}{G} + mS^2 + nS\right)}_{A(G)} \\ & \cdot \underbrace{\left(\frac{1}{N^\alpha} + \frac{k}{N_a^\alpha} + h\frac{N_a}{N}\right)}_{B} \\ & + \underbrace{\left(\frac{a}{N^\alpha} + \frac{c}{N_a^\alpha} + \frac{b}{D^\beta} + s\right)}_{C} \\ = & A(G) \cdot B + C \end{aligned} \tag{16}$$

Compute $\frac{\partial L}{\partial G}$ and set to 0:

$$\frac{\partial L}{\partial G} = \left(e - \frac{f}{G^2}\right) \cdot B = 0 \tag{17}$$

Since $B \neq 0$, we get $e - \frac{f}{G^2} = 0$. Therefore:

$$G^2 = \frac{f}{e} \implies G = \sqrt{\frac{f}{e}} \quad (G > 0) \tag{18}$$

The second derivative checks to confirm a minimum.

$$\frac{\partial^2 L}{\partial G^2} = \frac{2f}{G^3} \cdot B > 0 \tag{19}$$

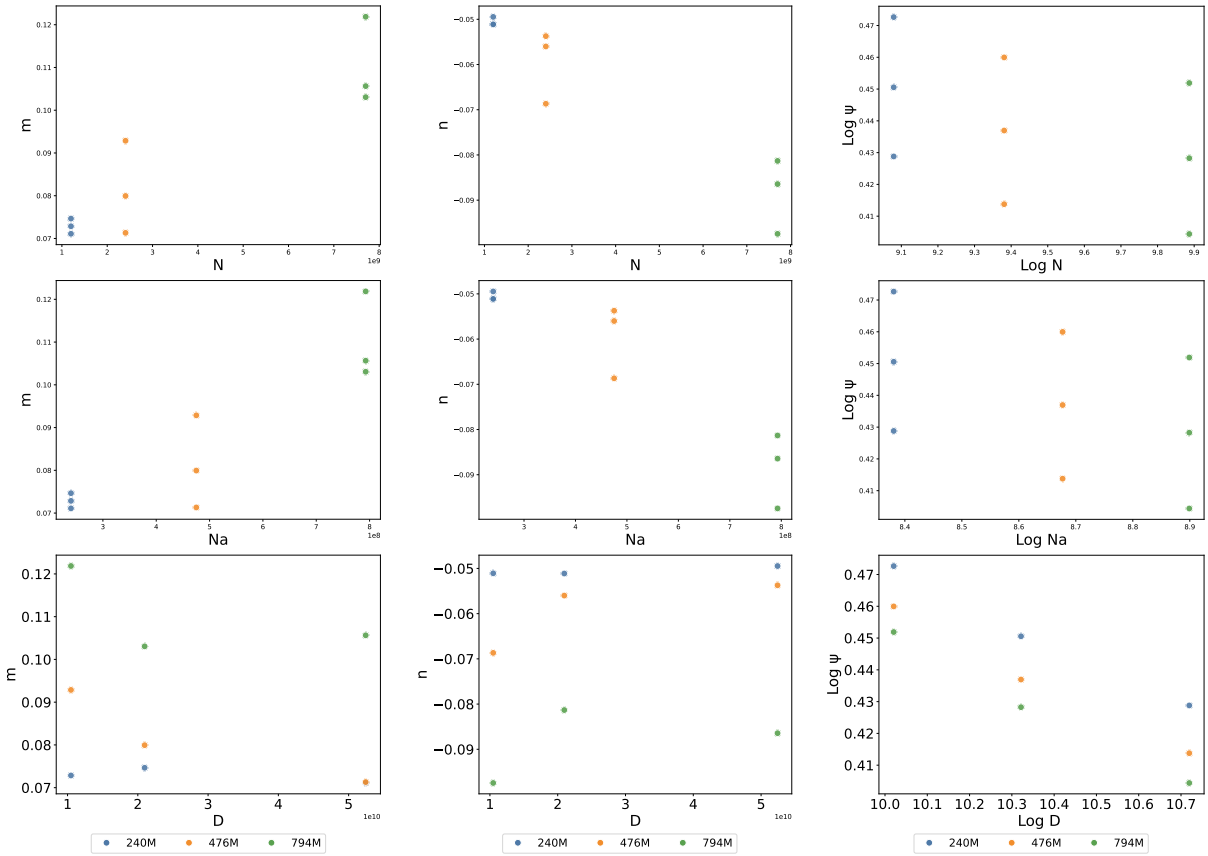


Figure 12: Correlations between m, n, ψ in Eq. 11 and N, N_a, D. While, m and n are independent of D, and both vary monotonically with N and N_a, while ψ follows a power-law relationship with N, N_a, and D.

The extreme point (minimum) is:

$$G_{opt} = \sqrt{\frac{f}{e}} \tag{20}$$

Substituting the optimized hyperparameters into the above Eq. 20 yields $G_{\text{opt}} \approx 6.78$.

To substantiate the validity of our conclusions, we conducted an analysis on the configurations of mainstream industrial MoE models, with detailed specifications in Table 3.

Theoretically, the optimal granularity is $G \approx 6.78$. In practice, however, G is typically constrained to infrastructure- and hardware-friendly values. Building upon our joint MoE scaling law, we derive practical ranges for G tailored to mainstream MoE models, with the loss increase constrained to within 0.001.

As shown in Table 3, the practical ranges of G predominantly fall within $[4, 12]$ across diverse model configurations. Notably, both the theoretically optimal $G \approx 7$ and the derived practical ranges align closely with configurations adopted by mainstream MoE models at varying scales, providing implicit validation of our scaling law insights regarding G.

E Derivation of Theoretical Optimal S

Analogous to the derivation of the optimal value of G, the isolation of S from other factors followed by the computation of its first-order derivative $\frac{\partial L}{\partial S}$—with the derivative set to 0:

$$\frac{\partial L}{\partial S} = (2mS + n) \cdot B = 0 \tag{21}$$

Since $B \neq 0$, we get $2mS + n = 0$. Therefore:

$$S_{opt} = -\frac{n}{2m} \tag{22}$$

Likewise, the second-derivative test confirms that S_{opt} corresponds to a minimum value. Substituting the optimized hyperparameters into the above Eq. 22 yields $S_{\text{opt}} \approx 0.31$.

The presence of shared experts yields substantial loss reduction. As S increases further, the loss continues to decrease marginally before rising sharply beyond the optimal point, with this effect becoming increasingly pronounced as N scales up in Figure 13. In summary, our findings lead

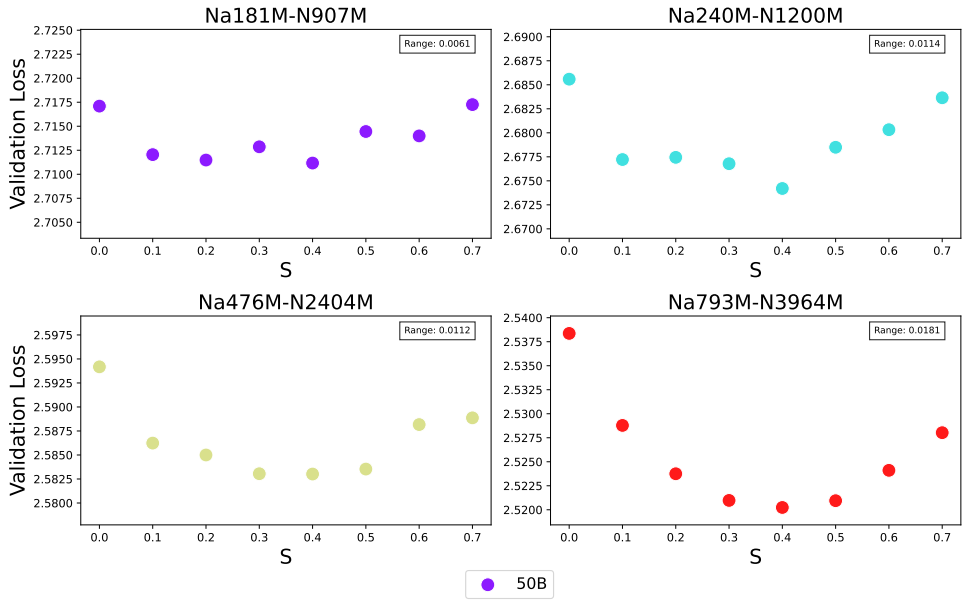


Figure 13: Illustration of the scaling law of S becoming increasingly prominent with growing model size.

Table 3: Analyses of mainstream MoE models. Based on our scaling law, the theoretically optimal granularity is $G \approx 6.78$, close with configurations adopted by industrial LLMs. The practical range indicates recommended settings of G with negligible performance loss (≤ 0.001).

Model	G (n_s+TopK)	G (Actual)	G (Practical Range, Thr $= 0.001$)
Hunyuan-A13B	$1+8$	9	$[4.99, 9.21]$
DeepSeek-V4-Flash (DeepSeek-AI, 2026)	$1+6$	7	$[4.19, 10.98]$
Hy3 preview	$1+8$	9	$[4.43, 10.38]$
Deepseek-V3 (Liu et al., 2024)	$1+8$	9	$[4.20, 10.93]$
GLM-5.1 (Zeng et al., 2026)	$1+8$	9	$[4.18, 11.00]$
Kimi-K2.6 (Team et al., 2026)	$1+8$	9	$[3.85, 11.95]$
Ling-2.6-1T	$1+8$	9	$[3.76, 12.21]$
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	$1+6$	7	$[3.85, 11.95]$

to the following recommendation: shared experts are essential, and for the optimal activated expert number $G = 7$, setting 1 or 2 shared experts is advisable. This conclusion is almost consistent with DeepSeek (DeepSeek-AI, 2026; Liu et al., 2024).

F Derivation of Theoretical and Practical Optimal N_a/N

Theoretically Analysis. Eq. 12 contains two types of terms involving N_a: one where N_a appears in the denominator and one where it appears in the numerator. Since these two terms exert opposing effects on the loss, an optimal N_a exists for any given configuration of the remaining factors (e.g., total model size N). This has been experimentally validated on both the Torchtitan (Liang et al., 2024) and Megatron (Shoeybi et al., 2019) pre-training frameworks. To derive the optimal ratio $\left(\frac{N_a}{N}\right)_{\text{opt}}$, we let $r = \frac{N_a}{N}$ and decompose L as follows:

$$L_{N,D,G,S}(r) = \underbrace{\left(eG + \frac{f}{G} + mS^2 + nS\right)}_{A} \cdot \underbrace{\left(\frac{1}{N^\alpha} + \frac{k}{(rN)^\alpha} + hr\right)}_{B(r)} + \underbrace{\left(\frac{a}{N^\alpha} + \frac{c}{(rN)^\alpha} + \frac{b}{D^\beta} + s\right)}_{C(r)} \tag{23}$$

Compute first derivative $\frac{\partial L}{\partial r}$ and set to 0:

$$\frac{\partial L}{\partial r} = A \cdot \frac{\partial B}{\partial r} + \frac{\partial C}{\partial r} = 0 \tag{24}$$

Calculate partial derivatives:

$$\frac{\partial B}{\partial r} = -\frac{\alpha k}{N^\alpha r^{\alpha+1}} + h, \tag{25}$$

$$\frac{\partial C}{\partial r} = -\frac{\alpha c}{N^\alpha r^{\alpha+1}} \tag{26}$$

Substitute and simplify:

$$A\left(-\frac{\alpha k}{N^{\alpha}r^{\alpha+1}}+h\right)-\frac{\alpha c}{N^{\alpha}r^{\alpha+1}}=0 \tag{27}$$

Rearrange terms to isolate r:

$$Ah=\frac{\alpha}{N^{\alpha}r^{\alpha+1}}(Ak+c) \implies r^{\alpha+1}=\frac{\alpha(Ak+c)}{AhN^{\alpha}} \tag{28}$$

Thus:

$$r=\left(\frac{\alpha(Ak+c)}{AhN^{\alpha}}\right)^{\frac{1}{\alpha+1}} \tag{29}$$

For $\alpha>0$, $\frac{\partial^2 L}{\partial r^2}=\frac{\alpha(\alpha+1)(Ak+c)}{N^{\alpha}r^{\alpha+2}}>0$, confirming a minimum.

Therefore, the optimal point of $(\frac{N_a}{N})_{opt_t}$ is:

$$\left(\frac{N_a}{N}\right)_{opt_t}=\left(\frac{\alpha\cdot[k\cdot\Phi(G,S)+c]}{hN^{\alpha}\cdot\Phi(G,S)}\right)^{\frac{1}{\alpha+1}} \tag{30}$$

From the expression of $\left(\frac{N_a}{N}\right)_{\text{opt}}$, the theoretically optimal activation ratio is jointly determined by G, S, and N. As established in the foregoing analysis, G (Appendix D) and S (Appendix E) each admit independent optimal values, which remain fixed regardless of model scale. Consequently, $\left(\frac{N_a}{N}\right)_{\text{opt}}$ decreases monotonically as N increases, indicating that the optimal MoE architecture becomes sparser with larger total model size. This trend is consistent with the design choices of current industry-level MoE models (Team et al., 2026; Zeng et al., 2026; DeepSeek-AI, 2026).

Practical Efficiency-aware Analysis. However, the theoretically optimal activation ratio $\frac{N_a}{N}$ derived in Eq. 30 cannot be directly applied to real-world MoE architecture design, as computational efficiency is an equally critical consideration. Specifically, as N_a approaches its theoretical optimum, performance gains become increasingly marginal while the associated costs rise steadily. It is therefore necessary to identify a practical, efficiency-aware optimal $\frac{N_a}{N}$ that balances performance gain against efficiency cost.

To this end, we define a loss gain threshold ΔLoss with a step size of $\Delta N_a = 0.01N$. As N_a is incrementally increased by this step size, the marginal loss reduction eventually falls below ΔLoss, at which point we define the model as having reached the practical efficiency-aware optimal $\frac{N_a}{N}$. The detailed derivation proceeds as follows:

Hence, for a given model size N, our MoE scaling law yields a practically applicable range for N_a, i.e., spanning the interval from the practical efficiency-aware optimal point to the theoretically optimal point, i.e., $N_a \in [\left(\frac{N_a}{N}\right)_{opt_e}, \left(\frac{N_a}{N}\right)_{opt_t}]$.

Algorithm 1: Find Efficiency-aware Optimal N_a

```
Function FindEfficiencyAwareNa(N, D, G, S, threshold):
    step = 0.01 × N;
    N_a_prev ← step;
    loss_prev ← L(N, D, N_a_prev, G, S);
    iteration = 1;
    max_iterations = n;
    while iteration ≤ max_iterations do
        N_a_current ← N_a_prev + step;
        loss_current ←
          L(N, D, N_a_current, G, S);
        loss_reduction ←
          loss_prev − loss_current;
        if loss_reduction < threshold then
            return N_a_current;
        N_a_prev ← N_a_current;
        loss_prev ← loss_current;
        iteration ← iteration + 1;
    return None;
```

To validate this conclusion, we analyze the configurations of mainstream industry-level MoE models, with detailed specifications provided in Table 4. The results show that the activated model sizes N_a of most mainstream MoE models fall within our recommended ranges, demonstrating that practical MoE designs can benefit from our scaling law in jointly optimizing both effectiveness and efficiency.

Discussion on the Optimal G and N_a. We clarify a potential misconception: the existence of optimal values for G and N_a in our MoE scaling law may appear to conflict with the intuition that larger G and N_a generally lead to better performance. This apparent contradiction, however, dissolves upon closer inspection. In practice, the most natural way to increase N_a while keeping N fixed is to increase G simultaneously, causing both N_a and G to grow in tandem. Under this coupled adjustment, our scaling law correctly predicts that the loss decreases monotonically, consistent with the intuition that more activated parameters improve performance. The non-monotonic behavior emerges only when a single factor is varied in isolation. When G alone is increased with all other factors fixed, the number of routed experts grows

Table 4: Theoretical and practical efficiency-aware optimal N_a/N analysis for mainstream MoE models. Although the theoretically optimal N_a/N serves as an upper bound, a wide range of N_a values below it yield only marginal loss gains, making the practical efficiency-aware optimal N_a/N a more actionable guide for real-world MoE design. Most mainstream models fall within our recommended practical range. A few recent models adopt notably lower activation ratios, likely reflecting commercial deployment considerations and a deliberate trade-off between efficiency and performance.

Model	N_a-N **(Actual)**	N_a/N **Theoretical Opt**	N_a/N **Practical Opt** (Δ**Loss** ≤ 0.005)
Hunyuan-A13B	13B-80B	33.16% (26.5B)	6.00% (5.6B)
DeepSeek-V4-Flash (DeepSeek-AI, 2026)	13B-284B	25.99% (73.8B)	6.00% (17.0B)
Hy3 preview	21B-295B	25.80% (75.1B)	6.00% (17.7B)
Deepseek-V3.1 (Liu et al., 2024)	37B-671B	22.02% (147.8B)	5.00% (33.6B)
GLM-5.1 (Zeng et al., 2026)	40B-744B	21.59% (160.6B)	5.00% (37.2B)
Kimi-K2.6 (Team et al., 2026)	32B-1T	20.40% (204.0B)	5.00% (50.0B)
Ling-2.6-1T	63B-1T	20.40% (204.0B)	5.00% (50.0B)
DeepSeek-V4-Pro (DeepSeek-AI, 2026)	49B-1.6T	18.63% (298.1B)	5.00% (80.0B)

while each expert's dimension shrinks proportionally. Although finer-grained expert partitioning can enhance performance up to a point, exceeding the optimal granularity eventually impairs model performance. Similarly, when N_a alone is increased with other factors fixed, the expert dimensions grow but the total number of experts decreases, undermining the sparsity advantage of MoE. This leads to structural distortion in the MoE architecture, disrupting the combinatorial activation mechanism and degrading performance.

G Compute-Optimality with Fixed Configurations

We controll the total computation cost $C = D \cdot N_a$ and analyze the relationship between the optimal loss and C. According to our Implications #1 and #2, G and S have optimal values. Thus, the term $\Phi = eG + \frac{f}{G} + mS^2 + nS$ can be expressed as a constant term $const$ under the optimal configuration. When N is fixed, substituting $D = \frac{C}{N_a}$ into our joint MoE scaling law (Eq. 12) yields the following result:

$$L(N_a, C) = C_0 + (\Phi \cdot k + c)\frac{1}{N_a^{\alpha}} + \frac{\Phi \cdot h}{N} \cdot N_a + \frac{bN_a^{\beta}}{C^{\beta}}, \tag{31}$$

where $\Phi = eG + \frac{f}{G} + mS^2 + nS$ and $C_0 = \frac{\Phi + a}{N^{\alpha}} + s$. To find the optimal N_a (denoted N_a^*) that minimizes $L(N_a)$, we take the first derivative with respect to N_a and set $\frac{dL}{dN_a} = 0$:

$$\frac{b\beta {N_a^*}^{\beta-1}}{C^{\beta}} = \frac{\alpha(\Phi k + c)}{{N_a^*}^{\alpha+1}} - \frac{\Phi h}{N} \tag{32}$$

Substituting N_a^* into Eq. 31 yields the optimal loss L^*:

$$L^*(C) = C_0 + \frac{(\Phi k + c)(\alpha + \beta)}{\beta} {N_a^*}^{-\alpha} + \frac{\Phi h(\beta - 1)}{N\beta} N_a^*$$
$$\text{s.t.} \quad \frac{b\beta {N_a^*}^{\beta-1}}{C^{\beta}} = \frac{\alpha(\Phi k + c)}{{N_a^*}^{\alpha+1}} - \frac{\Phi h}{N} \tag{33}$$

To facilitate understanding, we conduct the derivation under the predefined fixed configuration: $G = 7$, $S = 0.31$ and $N = 1\text{T}$. The expression for $L^*(C)$ is provided as follows and illustrated in Figure 14:

$$L^*(C) \approx 1.87 + 576 \cdot C^{-0.158} \tag{34}$$

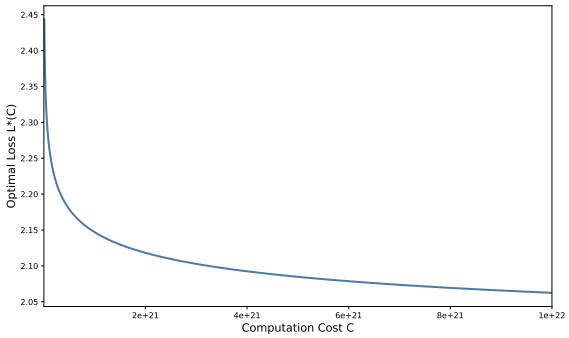


Figure 14: Illustration of $L^*(C)$ under the configuration of $N = 1\text{T}$, $G = 7$ and $S = 0.31$.

H Derivation of the u-v Relationship For N_a

To isolate N_a as the sole varying factor, we note from Eq. 4 that with G and S fixed, N_a can be controlled by jointly adjusting the expert dimension and the number of routed experts. Specifically, we scale the expert dimension by a factor of u and the number of routed experts by a factor of v. To ensure that N remains constant under this adjustment, the following constraint must hold:

$$\begin{aligned}&\big(4d_{\text{head}} \cdot n_h + 3d_{\text{expert}}(SG + n_e)\big)d_{\text{hidden}} \cdot l \\ =&\big(4d_{\text{head}} \cdot n_h + 3d_{\text{expert}} \cdot u(SG + n_e \cdot v)\big)d_{\text{hidden}} \cdot l\end{aligned} \tag{35}$$

Through formula transformation, we can derive:

$$v = \frac{(1-u) \cdot S \cdot G + n_e}{u \cdot n_e} \tag{36}$$

Thus, by setting the ratio of u to v according to the aforementioned Eq. 36, the controlled variation of N_a can be achieved. For instance, when $N = 2.4$B, $G = 20$ and $S = 0.2$, N_a takes values in the set $\{303, 476, 819, 1507, 2196\}$M, with the corresponding d_{expert} being $\{112, 224, 448, 896, 1344\}$ and the respective n_e being $\{260, 128, 62, 29, 18\}$.

I Related Work

I.1 MoE Architecture

In language models, MoE enables experts to learn diverse knowledge and combine their outputs (Abdin et al., 2024; Team et al., 2025; Zeng et al., 2025; Lieber et al., 2024). (Shazeer et al., 2017) expanded upon this with the Sparsely-Gated Mixture-of-Experts (SMoE) layer and Top-K routing, which selects a fixed number of experts for each token. This was further developed by Gshard (Lepikhin et al., 2020) and SwitchTransformer (Fedus et al., 2022) by integrating MoE into Transformer feedforward layers with Top-1 and Top-2 routing. More recently, (Dai et al., 2024) proposed modifying the MoE layer by subdividing experts into smaller experts and adding shared experts into the architecture. These advancements continue to enhance the efficiency and flexibility of MoE.

Meanwhile, recent studies have scaled MoE to larger model sizes and a greater number of experts, as demonstrated by Kimi K2 (Kimi et al., 2025), DeepSeek-V3 (Liu et al., 2024), and Mixture of a Million Experts (He, 2024). In parallel, several works have explored alternative routing strategies and expert designs, including heterogeneous experts in HMoE (Wang et al., 2024a), autonomous expert activation in AoE (Lv et al., 2025), and probabilistic Top-P routing (Zhou et al., 2022). In this work, we adopt the classical Top-K routing strategy as our main setting.

I.2 Scaling Laws of LLMs

Scaling laws for LLMs describe how model performance depends on factors such as model size and training data. In dense transformers, (Kaplan et al., 2020) first studied scaling laws and showed that the final model perplexity follows a power-law relationship with both model size and data size. Building on this, (Hoffmann et al., 2022) extended the analysis by incorporating variable cosine cycle lengths and proposed a revised scaling formulation. Scaling behavior has also been examined in alternative architectures, particularly in MoE models. (Clark et al., 2022) investigated MoE scaling laws under a fixed dataset, focusing on the impact of model size and the number of experts. (Krajewski et al., 2024) studied how scaling changes with different levels of expert granularity in MoE architectures. (Abnar et al., 2025) analyzed scaling laws with respect to total model size N, dataset size D and the fraction of inactive experts, while (Ludziejewski et al., 2025) examined the joint effects of multiple factors, including the activated model size N_a, data size D. Detailed comparison in Table 5.

Beyond architectures, Step Law (Li et al., 2025b) and relevant studies (Shuai et al., 2024; Zhang et al., 2024; McCandlish et al., 2018) provide principles for learning rate, weight decay and batch size, while Farseer (Li et al., 2025c) refines loss scaling for accurate extrapolation. Other works include parallel scaling (Chen et al., 2025) for improving compute efficiency and SynthLLM (Qin et al., 2025) for analyzing the scalability of synthetic data. More recently, (Sun et al., 2025) proposed a joint scaling law for floating point quantization training, highlighting the influence of exponent and mantissa bits, the effect of critical data size and the optimal precision range for efficient LLM training. Overall, these works extend scaling laws across optimizers, architectures and data, providing guidance for LLM design and training. Nevertheless, the community is urgent to build a more comprehensive MoE scaling law.

We selected two related scaling laws for comparison, each focusing on different aspects. Below, we briefly introduce their main ideas and describe

Table 5: Comparison with existing MoE scaling laws.

	N	Na	D	G	S
Scaling Laws for Fine-Grained MoE (Krajewski et al., 2024)	✓	✗	✓	✓	✗
Parameters vs. FLOPs (Abnar et al., 2025)	✓	✓	✓	✗	✗
Joint MoE Scaling Laws (Ludziejewski et al., 2025)	✓	✓	✓	✗	✗
Scaling Laws Across Architectures (Wang et al., 2024b)	✓	✗	✓	✗	✗
Unified Routed LMs (Clark et al., 2022)	✓	✗	✗	✗	✗
Our MoE Scaling Law	✓	✓	✓	✓	✓

how we use them to compare with the scaling law derived from our experiments.

I.2.1 Scaling Laws for Fine-Grained Mixture-of-Experts

(Krajewski et al., 2024) introduced a *granularity* factor into MoE scaling laws. They model the training loss $\mathcal{L}$ as a function of the number of **active parameters** N, the dataset size D and the **granularity** G:

$$\mathcal{L}(N, D, G) = c + \left(\frac{g}{G^{\gamma}} + a\right)\frac{1}{N^{\alpha}} + \frac{b}{D^{\beta}} \tag{37}$$

Here c is the irreducible loss. The term $N^{-\alpha}$ captures the effect of model size, adjusted by G: finer experts (larger G) reduce this contribution. The last term $D^{-\beta}$ reflects the improvement from more data. Overall, the law indicates diminishing returns from N and D, with better performance at higher G. As shown in Figure 5, this fitted result is less accurate than ours. The detailed fitted constants and values in Eq. 37 are presented below in Table 6.

Table 6: Fitted constants and their values in Eq. 37.

Constant	Value
a	81.9404
α	0.2380
b	3195.7723
β	0.3695
c	1.7709
g	0.0004
γ	0.0028

I.2.2 Scaling Laws for Optimal Sparsity in MoE

(Abnar et al., 2025) analyzed the effect of **sparsity** S, defined as the fraction of inactive experts. Their scaling law is:

$$L(N, D, S) = \frac{a}{N^{\alpha}} + \frac{b}{D^{\beta}} + \frac{c}{(1-S)^{\lambda}} + \frac{d}{(1-S)^{\delta}N^{\gamma}} + e \tag{38}$$

The first two terms represent the standard effects of model and data size. The third term penalizes high sparsity, while the fourth couples sparsity with model size. The constant e is an offset. This form shows that both the level of sparsity and its interaction with N influence the loss. As shown in Figure 5, our method achieves a better fit than this result. The detailed fitted constants and values in Eq. 38 are presented below in Table 7.

Table 7: Fitted constants and their values in Eq. 38.

Constant	Value
a	43.2889
α	0.1948
b	8280.7176
β	0.4145
e	1.6351
c	0.0003
d	9.4982
λ	1.8469
δ	0.0802
γ	14.0591

I.2.3 Comparison Methodology

For a fair comparison, all baseline scaling laws are fitted using the same dataset as ours, with parameters optimized on identical experimental results under the same settings, ensuring that any performance differences arise solely from differences in functional form. We additionally fit other MoE scaling laws (Ludziejewski et al., 2025; Clark et al., 2022), which likewise underperform relative to ours. This is likely because our formulation encompasses a more comprehensive set of factors (D,

N, N_a, G, S) over a broader range of MoE configurations, while carefully characterizing marginal effects and invoking Occam's razor to keep both the hyperparameters and functional forms parsimonious.

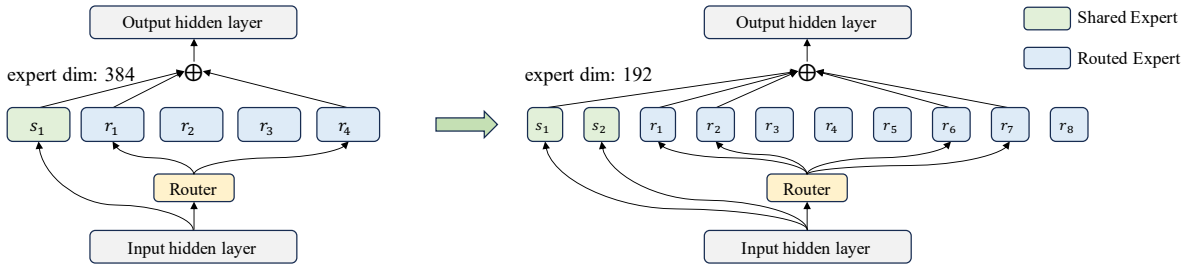
J Variation of different factors

In our study, the core factors, including N_a, N, G, and S, are varied in strict adherence to the principle of controlling variables. We have already discussed the methods used to control these factors during the experiments in Appendix C. To clarify the experimental procedure, we provide a series of illustrations in Figure 15 that demonstrate how each variable is systematically adjusted:

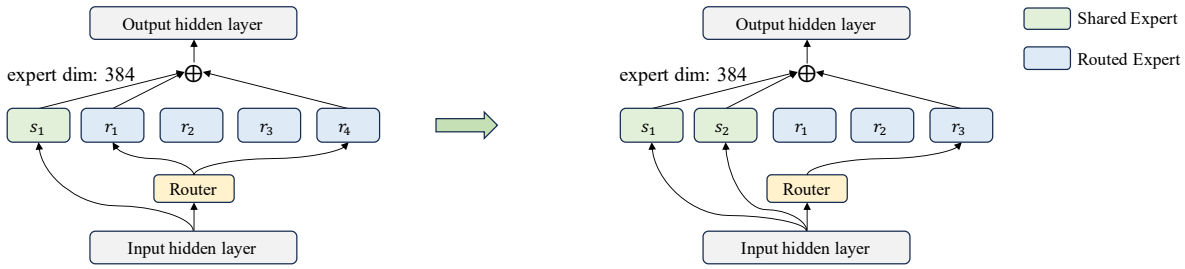
K Detailed Settings of Experiments

We show the detailed configurations of our experiments as follows. Precisely, the value ranges of various factors in the experimental setup are as follows: for fitting data points, $G \in (1, 20]$, $S \in [0.0, 0.8]$, $N_a \in [30\text{M}, 2.2\text{B}]$, $N \in [133\text{M}, 3.4\text{B}]$ and $D \in [10\text{B}, 50\text{B}]$; for validation data points, $G \in (1, 40]$, $S \in [0.0, 0.7]$, $N_a \in [453\text{M}, 6.6\text{B}]$, $N \in [2.4\text{B}, 30\text{B}]$ and $D \in [10\text{B}, 100\text{B}]$. Notably, to mitigate the impact of random seeds, our fitting data points include experimental data points obtained from identical configurations with different random seeds. It is worth noting that none of the validation data points are involved in hyperparameter fitting or the selection of the optimal fitted results. Instead, they are reserved exclusively for validation to ensure an unbiased evaluation of the scaling laws.

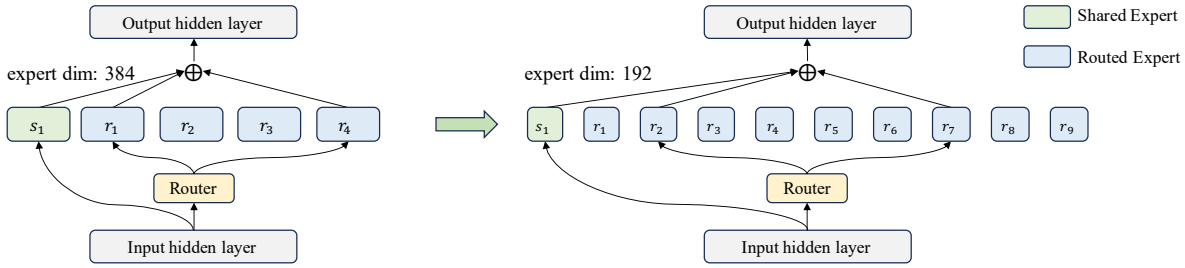
We should highlight that our joint MoE scaling law is designed to accommodate practical experimental configurations. Therefore, the MoE configurations are more focused on the relatively practical settings, with partial of extreme settings to reveal the marginal effect of different factors. For the validation data, we adopt larger model/data sizes and broader ranges of factors to evaluate the effectiveness of our scaling law. For some uncommon and impractical experimental setups, such as those with extremely large G and exceptionally high N_a/N ratios simultaneously (whose efficiency is unsatisfactory in the view of MoE models), perfect fitting is not pursued by our MoE scaling law. The detailed configurations are in Table 8.



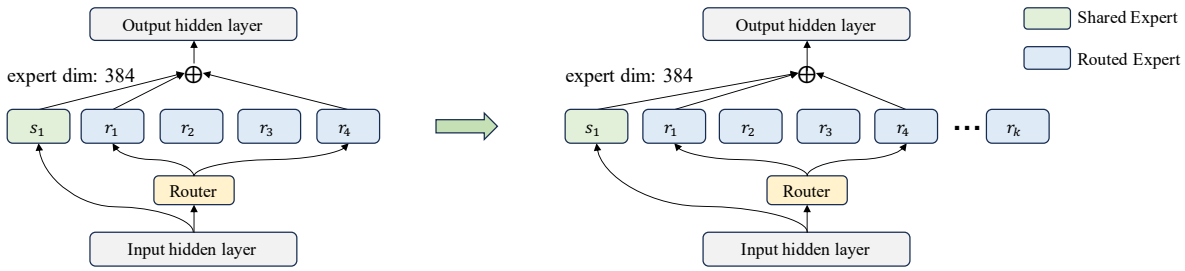
(a) Illustration of the variation of G.



(b) Illustration of the variation of S.



(c) Illustration of the variation of N_a.



(d) Illustration of the variation of N.

Figure 15: Illustration of the variation of factor G, S, N_a and N under controlled conditions.

Table 8: All configurations of experiments. The last column is **Label**, where ✓ indicates that the data point is used for validation and ✗ indicates that the data point is used for fitting or for observing marginal patterns.

	Na	N	D	G	S	Label		Na	N	D	G	S	Label
0	48M	247M	10B	2.5	0.2	✗	1	48M	247M	10B	5	0.2	✗
2	48M	247M	10B	10	0.2	✗	3	48M	247M	10B	15	0.2	✗
4	48M	247M	10B	20	0.2	✗	5	48M	247M	20B	2.5	0.2	✗
6	48M	247M	20B	5	0.2	✗	7	48M	247M	20B	10	0.2	✗
8	48M	247M	20B	15	0.2	✗	9	48M	247M	20B	20	0.2	✗
10	48M	247M	50B	2.5	0.2	✗	11	48M	247M	50B	5	0.2	✗
12	48M	247M	50B	10	0.2	✗	13	48M	247M	50B	15	0.2	✗
14	48M	247M	50B	20	0.2	✗	15	99M	496M	10B	2.5	0.2	✗
16	99M	496M	10B	5	0.2	✗	17	99M	496M	10B	10	0.2	✗
18	99M	496M	10B	15	0.2	✗	19	99M	496M	10B	20	0.2	✗
20	99M	496M	20B	2.5	0.2	✗	21	99M	496M	20B	5	0.2	✗
22	99M	496M	20B	10	0.2	✗	23	99M	496M	20B	15	0.2	✗
24	99M	496M	20B	20	0.2	✗	25	99M	496M	50B	2.5	0.2	✗
26	99M	496M	50B	5	0.2	✗	27	99M	496M	50B	10	0.2	✗
28	99M	496M	50B	15	0.2	✗	29	99M	496M	50B	20	0.2	✗
30	181M	907M	10B	2.5	0.2	✗	31	181M	907M	10B	5	0.2	✗
32	181M	907M	10B	10	0.2	✗	33	181M	907M	10B	15	0.2	✗
34	181M	907M	10B	20	0.2	✗	35	181M	907M	20B	2.5	0.2	✗
36	181M	907M	20B	5	0.2	✗	37	181M	907M	20B	10	0.2	✗
38	181M	907M	20B	15	0.2	✗	39	181M	907M	20B	20	0.2	✗
40	181M	907M	50B	2.5	0.2	✗	41	181M	907M	50B	5	0.2	✗
42	181M	907M	50B	10	0.2	✗	43	181M	907M	50B	15	0.2	✗
44	181M	907M	50B	20	0.2	✗	45	48M	133M	10B	10	0.2	✗
46	48M	247M	10B	10	0.2	✗	47	48M	473M	10B	10	0.2	✗
48	48M	926M	10B	10	0.2	✗	49	48M	133M	20B	10	0.2	✗
50	48M	247M	20B	10	0.2	✗	51	48M	473M	20B	10	0.2	✗
52	48M	926M	20B	10	0.2	✗	53	48M	133M	50B	10	0.2	✗
54	48M	247M	50B	10	0.2	✗	55	48M	473M	50B	10	0.2	✗
56	48M	926M	50B	10	0.2	✗	57	99M	269M	10B	10	0.2	✗
58	99M	496M	10B	10	0.2	✗	59	99M	949M	10B	10	0.2	✗
60	99M	1855M	10B	10	0.2	✗	61	99M	269M	20B	10	0.2	✗
62	99M	496M	20B	10	0.2	✗	63	99M	949M	20B	10	0.2	✗
64	99M	1855M	20B	10	0.2	✗	65	99M	269M	50B	10	0.2	✗
66	99M	496M	50B	10	0.2	✗	67	99M	949M	50B	10	0.2	✗
68	99M	1855M	50B	10	0.2	✗	69	181M	492M	10B	10	0.2	✗
70	181M	907M	10B	10	0.2	✗	71	181M	1738M	10B	10	0.2	✗
72	181M	3399M	10B	10	0.2	✗	73	181M	492M	20B	10	0.2	✗
74	181M	907M	20B	10	0.2	✗	75	181M	1738M	20B	10	0.2	✗
76	181M	3399M	20B	10	0.2	✗	77	181M	492M	50B	10	0.2	✗
78	181M	907M	50B	10	0.2	✗	79	181M	1738M	50B	10	0.2	✗
80	181M	3399M	50B	10	0.2	✗	81	48M	247M	10B	10	0.0	✗
82	48M	247M	10B	10	0.1	✗	83	48M	247M	10B	10	0.2	✗
84	48M	247M	10B	10	0.4	✗	85	48M	247M	20B	10	0.0	✗
86	48M	247M	20B	10	0.1	✗	87	48M	247M	20B	10	0.2	✗
88	48M	247M	20B	10	0.4	✗	89	48M	247M	50B	10	0.0	✗
90	48M	247M	50B	10	0.1	✗	91	48M	247M	50B	10	0.2	✗

Continued on next page

Table 8 continued from previous page

	Na	N	D	G	S	Label		Na	N	D	G	S	Label
92	48M	247M	50B	10	0.4	✗	93	99M	496M	10B	10	0.0	✗
94	99M	496M	10B	10	0.1	✗	95	99M	496M	10B	10	0.2	✗
96	99M	496M	10B	10	0.4	✗	97	99M	496M	20B	10	0.0	✗
98	99M	496M	20B	10	0.1	✗	99	99M	496M	20B	10	0.2	✗
100	99M	496M	20B	10	0.4	✗	101	99M	496M	50B	10	0.0	✗
102	99M	496M	50B	10	0.1	✗	103	99M	496M	50B	10	0.2	✗
104	99M	496M	50B	10	0.4	✗	105	181M	907M	10B	10	0.0	✗
106	181M	907M	10B	10	0.1	✗	107	181M	907M	10B	10	0.2	✗
108	181M	907M	10B	10	0.3	✗	109	181M	907M	10B	10	0.4	✗
110	181M	907M	10B	10	0.5	✗	111	181M	907M	10B	10	0.6	✗
112	181M	907M	10B	10	0.7	✗	113	181M	907M	20B	10	0.0	✗
114	181M	907M	20B	10	0.1	✗	115	181M	907M	20B	10	0.2	✗
116	181M	907M	20B	10	0.3	✗	117	181M	907M	20B	10	0.4	✗
118	181M	907M	20B	10	0.5	✗	119	181M	907M	20B	10	0.6	✗
120	181M	907M	20B	10	0.7	✗	121	181M	907M	50B	10	0.0	✗
122	181M	907M	50B	10	0.1	✗	123	181M	907M	50B	10	0.2	✗
124	181M	907M	50B	10	0.3	✗	125	181M	907M	50B	10	0.4	✗
126	181M	907M	50B	10	0.5	✗	127	181M	907M	50B	10	0.6	✗
128	181M	907M	50B	10	0.7	✗	129	240M	1209M	10B	10	0.0	✗
130	240M	1209M	10B	10	0.1	✗	131	240M	1209M	10B	10	0.2	✗
132	240M	1209M	10B	10	0.3	✗	133	240M	1209M	10B	10	0.4	✗
134	240M	1209M	10B	10	0.5	✗	135	240M	1209M	10B	10	0.6	✗
136	240M	1209M	10B	10	0.7	✗	137	240M	1209M	20B	10	0.0	✗
138	240M	1209M	20B	10	0.1	✗	139	240M	1209M	20B	10	0.2	✗
140	240M	1209M	20B	10	0.3	✗	141	240M	1209M	20B	10	0.4	✗
142	240M	1209M	20B	10	0.5	✗	143	240M	1209M	20B	10	0.6	✗
144	240M	1209M	20B	10	0.7	✗	145	240M	1209M	50B	10	0.0	✗
146	240M	1209M	50B	10	0.1	✗	147	240M	1209M	50B	10	0.2	✗
148	240M	1209M	50B	10	0.3	✗	149	240M	1209M	50B	10	0.4	✗
150	240M	1209M	50B	10	0.5	✗	151	240M	1209M	50B	10	0.6	✗
152	240M	1209M	50B	10	0.7	✗	153	476M	2404M	10B	10	0.0	✗
154	476M	2404M	10B	10	0.1	✗	155	476M	2404M	10B	10	0.3	✗
156	476M	2404M	10B	10	0.4	✗	157	476M	2404M	10B	10	0.5	✗
158	476M	2404M	10B	10	0.6	✗	159	476M	2404M	10B	10	0.7	✗
160	476M	2404M	20B	10	0.0	✗	161	476M	2404M	20B	10	0.1	✗
162	476M	2404M	20B	10	0.3	✗	163	476M	2404M	20B	10	0.4	✗
164	476M	2404M	20B	10	0.5	✗	165	476M	2404M	20B	10	0.6	✗
166	476M	2404M	20B	10	0.7	✗	167	476M	2404M	50B	10	0.0	✗
168	476M	2404M	50B	10	0.1	✗	169	476M	2404M	50B	10	0.3	✗
170	476M	2404M	50B	10	0.4	✗	171	476M	2404M	50B	10	0.5	✗
172	476M	2404M	50B	10	0.6	✗	173	476M	2404M	50B	10	0.7	✗
174	240M	1209M	10B	5	0.0	✗	175	240M	1209M	10B	5	0.2	✗
176	240M	1209M	10B	5	0.4	✗	177	240M	1209M	10B	5	0.6	✗
178	240M	1209M	10B	5	0.8	✗	179	240M	1209M	20B	5	0.0	✗
180	240M	1209M	20B	5	0.2	✗	181	240M	1209M	20B	5	0.4	✗
182	240M	1209M	20B	5	0.6	✗	183	240M	1209M	20B	5	0.8	✗
184	240M	1209M	50B	5	0.0	✗	185	240M	1209M	50B	5	0.2	✗
186	240M	1209M	50B	5	0.4	✗	187	240M	1209M	50B	5	0.6	✗

Continued on next page

Table 8 continued from previous page

	Na	N	D	G	S	Label		Na	N	D	G	S	Label
188	240M	1209M	50B	5	0.8	✗	189	476M	2404M	10B	5	0.0	✗
190	476M	2404M	10B	5	0.2	✗	191	476M	2404M	10B	5	0.4	✗
192	476M	2404M	10B	5	0.6	✗	193	476M	2404M	10B	5	0.8	✗
194	476M	2404M	20B	5	0.0	✗	195	476M	2404M	20B	5	0.2	✗
196	476M	2404M	20B	5	0.4	✗	197	476M	2404M	20B	5	0.6	✗
198	476M	2404M	20B	5	0.8	✗	199	476M	2404M	50B	5	0.0	✗
200	476M	2404M	50B	5	0.2	✗	201	476M	2404M	50B	5	0.4	✗
202	476M	2404M	50B	5	0.6	✗	203	476M	2404M	50B	5	0.8	✗
204	31M	247M	10B	20	0.2	✗	205	31M	247M	20B	20	0.2	✗
206	31M	247M	30B	20	0.2	✗	207	31M	247M	50B	20	0.2	✗
208	64M	496M	10B	20	0.2	✗	209	99M	496M	10B	20	0.2	✗
210	170M	496M	10B	20	0.2	✗	211	312M	496M	10B	20	0.2	✗
212	453M	496M	10B	20	0.2	✗	213	64M	496M	20B	20	0.2	✗
214	99M	496M	20B	20	0.2	✗	215	170M	496M	20B	20	0.2	✗
216	312M	496M	20B	20	0.2	✗	217	453M	496M	20B	20	0.2	✗
218	64M	496M	30B	20	0.2	✗	219	99M	496M	30B	20	0.2	✗
220	170M	496M	30B	20	0.2	✗	221	312M	496M	30B	20	0.2	✗
222	453M	496M	30B	20	0.2	✗	223	64M	496M	50B	20	0.2	✗
224	99M	496M	50B	20	0.2	✗	225	170M	496M	50B	20	0.2	✗
226	312M	496M	50B	20	0.2	✗	227	453M	496M	50B	20	0.2	✗
228	116M	907M	10B	20	0.2	✗	229	181M	907M	10B	20	0.2	✗
230	310M	907M	10B	20	0.2	✗	231	570M	907M	10B	20	0.2	✗
232	829M	907M	10B	20	0.2	✗	233	116M	907M	20B	20	0.2	✗
234	181M	907M	20B	20	0.2	✗	235	310M	907M	20B	20	0.2	✗
236	570M	907M	20B	20	0.2	✗	237	829M	907M	20B	20	0.2	✗
238	116M	907M	30B	20	0.2	✗	239	181M	907M	30B	20	0.2	✗
240	310M	907M	30B	20	0.2	✗	241	570M	907M	30B	20	0.2	✗
242	829M	907M	30B	20	0.2	✗	243	116M	907M	50B	20	0.2	✗
244	181M	907M	50B	20	0.2	✗	245	310M	907M	50B	20	0.2	✗
246	570M	907M	50B	20	0.2	✗	247	829M	907M	50B	20	0.2	✗
248	304M	2404M	10B	20	0.2	✗	249	476M	2404M	10B	20	0.2	✗
250	820M	2404M	10B	20	0.2	✗	251	1508M	2404M	10B	20	0.2	✗
252	2196M	2404M	10B	20	0.2	✗	253	304M	2404M	20B	20	0.2	✗
254	476M	2404M	20B	20	0.2	✗	255	820M	2404M	20B	20	0.2	✗
256	1508M	2404M	20B	20	0.2	✗	257	2196M	2404M	20B	20	0.2	✗
258	304M	2404M	30B	20	0.2	✗	259	476M	2404M	30B	20	0.2	✗
260	820M	2404M	30B	20	0.2	✗	261	1508M	2404M	30B	20	0.2	✗
262	2196M	2404M	30B	20	0.2	✗	263	304M	2404M	50B	20	0.2	✗
264	476M	2404M	50B	20	0.2	✗	265	820M	2404M	50B	20	0.2	✗
266	1508M	2404M	50B	20	0.2	✗	267	2196M	2404M	50B	20	0.2	✗
268	22M	121M	10B	1	0.0	✗	269	22M	121M	10B	2	0.0	✗
270	22M	121M	10B	3	0.0	✗	271	22M	121M	10B	4	0.0	✗
272	22M	121M	10B	8	0.0	✗	273	22M	121M	10B	16	0.0	✗
274	22M	121M	20B	1	0.0	✗	275	22M	121M	20B	2	0.0	✗
276	22M	121M	20B	3	0.0	✗	277	22M	121M	20B	4	0.0	✗
278	22M	121M	20B	8	0.0	✗	279	22M	121M	20B	16	0.0	✗
280	22M	121M	50B	1	0.0	✗	281	22M	121M	50B	2	0.0	✗
282	22M	121M	50B	3	0.0	✗	283	22M	121M	50B	4	0.0	✗

Continued on next page

Table 8 continued from previous page

	Na	N	D	G	S	Label		Na	N	D	G	S	Label
284	22M	121M	50B	8	0.0	✗	285	22M	121M	50B	16	0.0	✗
286	30M	93M	10B	1	0.0	✗	287	30M	93M	10B	2	0.0	✗
288	30M	93M	10B	3	0.0	✗	289	30M	93M	10B	4	0.0	✗
290	30M	93M	10B	8	0.0	✗	291	30M	93M	10B	16	0.0	✗
292	30M	93M	20B	1	0.0	✗	293	30M	93M	20B	2	0.0	✗
294	30M	93M	20B	3	0.0	✗	295	30M	93M	20B	4	0.0	✗
296	30M	93M	20B	8	0.0	✗	297	30M	93M	20B	16	0.0	✗
298	30M	93M	50B	1	0.0	✗	299	30M	93M	50B	2	0.0	✗
300	30M	93M	50B	3	0.0	✗	301	30M	93M	50B	4	0.0	✗
302	30M	93M	50B	8	0.0	✗	303	30M	93M	50B	16	0.0	✗
304	30M	175M	10B	1	0.0	✗	305	30M	175M	10B	2	0.0	✗
306	30M	175M	10B	3	0.0	✗	307	30M	175M	10B	4	0.0	✗
308	30M	175M	10B	8	0.0	✗	309	30M	175M	10B	16	0.0	✗
310	30M	175M	20B	1	0.0	✗	311	30M	175M	20B	2	0.0	✗
312	30M	175M	20B	3	0.0	✗	313	30M	175M	20B	4	0.0	✗
314	30M	175M	20B	8	0.0	✗	315	30M	175M	20B	16	0.0	✗
316	30M	175M	50B	1	0.0	✗	317	30M	175M	50B	2	0.0	✗
318	30M	175M	50B	3	0.0	✗	319	30M	175M	50B	4	0.0	✗
320	30M	175M	50B	8	0.0	✗	321	30M	175M	50B	16	0.0	✗
322	30M	340M	10B	1	0.0	✗	323	30M	340M	10B	2	0.0	✗
324	30M	340M	10B	3	0.0	✗	325	30M	340M	10B	4	0.0	✗
326	30M	340M	10B	8	0.0	✗	327	30M	340M	10B	16	0.0	✗
328	30M	340M	20B	1	0.0	✗	329	30M	340M	20B	2	0.0	✗
330	30M	340M	20B	3	0.0	✗	331	30M	340M	20B	4	0.0	✗
332	30M	340M	20B	8	0.0	✗	333	30M	340M	20B	16	0.0	✗
334	30M	340M	50B	1	0.0	✗	335	30M	340M	50B	2	0.0	✗
336	30M	340M	50B	3	0.0	✗	337	30M	340M	50B	4	0.0	✗
338	30M	340M	50B	8	0.0	✗	339	30M	340M	50B	16	0.0	✗
340	41M	239M	10B	1	0.0	✗	341	41M	239M	10B	2	0.0	✗
342	41M	239M	10B	3	0.0	✗	343	41M	239M	10B	4	0.0	✗
344	41M	239M	10B	8	0.0	✗	345	41M	239M	10B	16	0.0	✗
346	41M	239M	20B	1	0.0	✗	347	41M	239M	20B	2	0.0	✗
348	41M	239M	20B	3	0.0	✗	349	41M	239M	20B	4	0.0	✗
350	41M	239M	20B	8	0.0	✗	351	41M	239M	20B	16	0.0	✗
352	41M	239M	50B	1	0.0	✗	353	41M	239M	50B	2	0.0	✗
354	41M	239M	50B	3	0.0	✗	355	41M	239M	50B	4	0.0	✗
356	41M	239M	50B	8	0.0	✗	357	41M	239M	50B	16	0.0	✗
358	476M	1301M	10B	10	0.2	✓	359	476M	2404M	10B	10	0.2	✓
360	476M	4604M	10B	10	0.2	✓	361	476M	9008M	10B	10	0.2	✓
362	476M	1301M	20B	10	0.2	✓	363	476M	2404M	20B	10	0.2	✓
364	476M	4604M	20B	10	0.2	✓	365	476M	9008M	20B	10	0.2	✓
366	476M	1301M	50B	10	0.2	✓	367	476M	2404M	50B	10	0.2	✓
368	476M	4604M	50B	10	0.2	✓	369	476M	9008M	50B	10	0.2	✓
370	793M	3964M	10B	10	0.0	✓	371	793M	3964M	10B	10	0.1	✓
372	793M	3964M	10B	10	0.2	✓	373	793M	3964M	10B	10	0.3	✓
374	793M	3964M	10B	10	0.4	✓	375	793M	3964M	10B	10	0.5	✓
376	793M	3964M	10B	10	0.6	✓	377	793M	3964M	10B	10	0.7	✓
378	793M	3964M	20B	10	0.0	✓	379	793M	3964M	20B	10	0.1	✓

Continued on next page

Table 8 continued from previous page

	Na	N	D	G	S	Label		Na	N	D	G	S	Label
380	793M	3964M	20B	10	0.2	✓	381	793M	3964M	20B	10	0.3	✓
382	793M	3964M	20B	10	0.4	✓	383	793M	3964M	20B	10	0.5	✓
384	793M	3964M	20B	10	0.6	✓	385	793M	3964M	20B	10	0.7	✓
386	793M	3964M	50B	10	0.0	✓	387	793M	3964M	50B	10	0.1	✓
388	793M	3964M	50B	10	0.2	✓	389	793M	3964M	50B	10	0.3	✓
390	793M	3964M	50B	10	0.4	✓	391	793M	3964M	50B	10	0.5	✓
392	793M	3964M	50B	10	0.6	✓	393	793M	3964M	50B	10	0.7	✓
394	1441M	7255M	10B	10	0.2	✓	395	1960M	7255M	10B	10	0.2	✓
396	2479M	7255M	10B	10	0.2	✓	397	3517M	7255M	10B	10	0.2	✓
398	6632M	7255M	10B	10	0.2	✓	399	1441M	7255M	20B	10	0.2	✓
400	1960M	7255M	20B	10	0.2	✓	401	2479M	7255M	20B	10	0.2	✓
402	3517M	7255M	20B	10	0.2	✓	403	6632M	7255M	20B	10	0.2	✓
404	1441M	7255M	50B	10	0.2	✓	405	1960M	7255M	50B	10	0.2	✓
406	2479M	7255M	50B	10	0.2	✓	407	3517M	7255M	50B	10	0.2	✓
408	6632M	7255M	50B	10	0.2	✓	409	453M	3964M	10B	2	0.5	✓
410	453M	3964M	10B	4	0.5	✓	411	453M	3964M	10B	6	0.5	✓
412	453M	3964M	10B	8	0.5	✓	413	453M	3964M	10B	10	0.5	✓
414	453M	3964M	10B	12	0.5	✓	415	453M	3964M	20B	2	0.5	✓
416	453M	3964M	20B	4	0.5	✓	417	453M	3964M	20B	6	0.5	✓
418	453M	3964M	20B	8	0.5	✓	419	453M	3964M	20B	10	0.5	✓
420	453M	3964M	20B	12	0.5	✓	421	453M	3964M	50B	2	0.5	✓
422	453M	3964M	50B	4	0.5	✓	423	453M	3964M	50B	6	0.5	✓
424	453M	3964M	50B	8	0.5	✓	425	453M	3964M	50B	10	0.5	✓
426	453M	3964M	50B	12	0.5	✓	427	566M	3964M	10B	3	0.33	✓
428	566M	3964M	10B	6	0.33	✓	429	566M	3964M	10B	9	0.33	✓
430	566M	3964M	10B	12	0.33	✓	431	566M	3964M	20B	3	0.33	✓
432	566M	3964M	20B	6	0.33	✓	433	566M	3964M	20B	9	0.33	✓
434	566M	3964M	20B	12	0.33	✓	435	566M	3964M	50B	3	0.33	✓
436	566M	3964M	50B	6	0.33	✓	437	566M	3964M	50B	9	0.33	✓
438	566M	3964M	50B	12	0.33	✓	439	476M	2404M	10B	2.5	0.2	✓
440	476M	2404M	20B	2.5	0.2	✓	441	476M	2404M	50B	2.5	0.2	✓
442	793M	3964M	100B	10	0.2	✓	443	476M	2404M	10B	20	0.2	✓
444	476M	2404M	20B	20	0.2	✓	445	476M	2404M	50B	20	0.2	✓
446	3070M	30249M	50B	20	0.2	✓	447	181M	907M	10B	40	0.2	✓
448	181M	907M	20B	40	0.2	✓	449	181M	907M	50B	40	0.2	✓

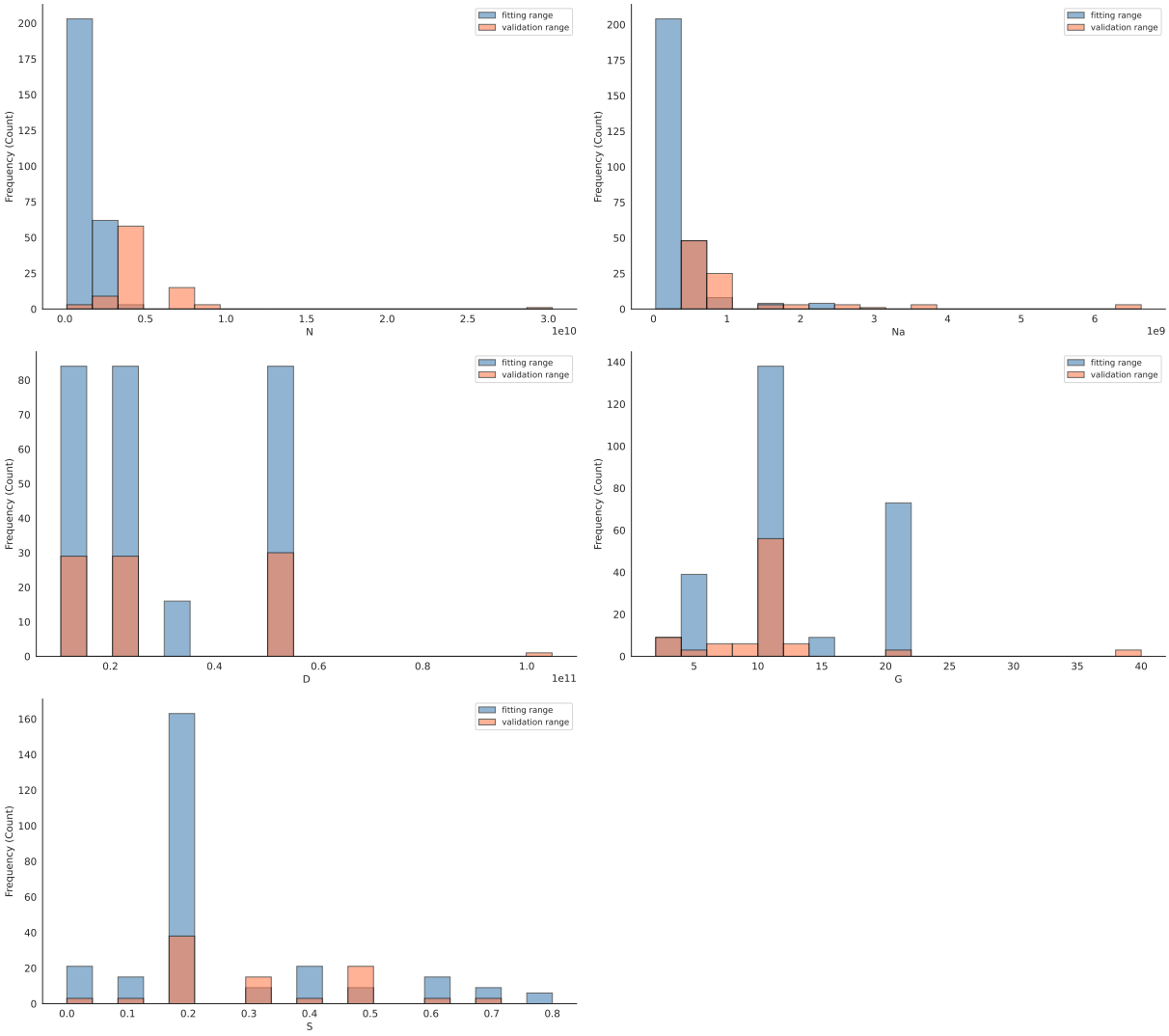


Figure 16: Distribution histograms of N, N_a, D, G, and S for fitting data and validation data.