DIVERSITY OF TRANSFORMER LAYERS: ONE ASPECT OF PARAMETER SCALING LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers improve predictably with model scale, yet the internal mechanisms underlying this behavior remain unclear. We study Transformer residual-stream components through bias–diversity decomposition, viewing their projected outputs as an ensemble. This perspective characterizes how component accuracy, diversity, and importance jointly affect prediction quality. We further develop an information-theoretic formulation in which adding informative layers monotonically improves prediction-error bounds. We characterize departures from diminishing returns by a dependence-increase term and show that the bounds exhibit approximate diminishing marginal improvements when this term is small. This provides a layer-level mechanism qualitatively consistent with the diminishing returns observed in parameter scaling laws. Experiments across eight NLP tasks and multiple LLM families show systematic relationships among bias, diversity, and accuracy. We also find that component reweighting alters the bias–diversity balance and that recurrent layer sharing substantially reduces measured diversity. Overall, our results establish layer diversity as a useful lens for understanding how Transformer components cooperate as model depth increases.

1 INTRODUCTION

Transformer (Vaswani et al., 2017) is one of the dominant architectures across a wide range of tasks, particularly as the backbone of large language models (LLMs). Their performance generally improves as model size increases, as described by parameter scaling laws (Hestness et al., 2017; Kaplan et al., 2020). These empirical laws characterize predictable performance improvements with scale, including capabilities such as in-context learning (Brown et al., 2020) and emergent behaviors in large models (Wei et al., 2022). However, it remains unclear how such model-level scaling behavior relates to the internal computations of Transformers.

Mechanistic interpretability has provided insights into the internal computations of Transformers and identified interpretable roles of attention and MLP modules (Geva et al., 2021; Olsson et al., 2022). In particular, studies of the residual stream (Elhage et al., 2021) show how embedding, Multi-head Attention (MHA), and Multi-layer Perceptron (MLP) outputs accumulate across layers. These observations suggest that individual layers contribute incrementally to the final prediction. Nevertheless, how the interactions among these layer-wise contributions relate to model scaling remains underexplored.

To investigate this connection, we analyze Transformer residual-stream components from the perspective of *diversity*. We interpret projected residual-stream outputs as members of an ensemble and apply bias-diversity decomposition (Krogh & Vedelsby, 1994). In this formulation, *bias* measures the error of individual component predictions relative to the ground truth, while *diversity* measures differences among their predictions. This perspective allows us to characterize how the quality and complementarity of residual-stream components jointly affect the final prediction. Our analysis yields the following observations:

- Lower bias and higher diversity jointly reduce the prediction error of the aggregated output (Theorem 1).
- Bias and diversity can be decomposed across residual-stream modules, while component magnitudes act as implicit weights affecting both quantities (Theorems 1, 2, and 3).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

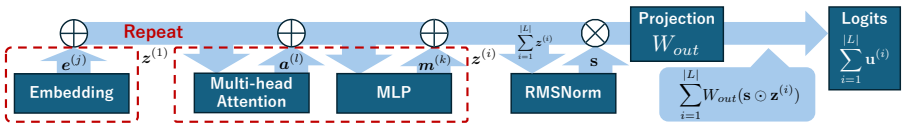

Figure 1: The overview of the residual stream in pre-layer normalization type Transformers.

- Bias and diversity are generally coupled, making it difficult to improve them independently in some cases (Theorems 4 and 8).

We further study how this perspective relates to increasing Transformer depth. Rather than deriving parameter scaling laws directly, we focus on whether additional layers exhibit a diminishing-return mechanism consistent with such laws. Using an information-theoretic formulation of ensemble diversity (Brown, 2009; Zhou & Li, 2010), we characterize the information jointly provided by layer outputs about the prediction target. This analysis gives the following results:

- The joint contribution of layer outputs can be characterized by *Information-Theoretic Diversity*, which reflects their relevance to the target and statistical dependence (Theorem 5).
- Adding layer outputs cannot worsen the corresponding information-theoretic prediction-error bounds because their joint mutual information with the target is non-decreasing (Theorem 6).
- We derive an exact identity for changes in marginal target information and show approximate diminishing returns when the resulting dependence-increase term is small (Theorem 7).

Finally, we evaluate the proposed perspective across multiple LLMs and eight NLP datasets: AG News (Zhang et al., 2015), ARC (Clark et al., 2018), BoolQ (Clark et al., 2019), MNLI (Williams et al., 2018), QQP (Wang et al., 2017), RTE, SST-2, and WIC (Wang et al., 2018). We analyze the relationships among prediction error, bias, diversity, and task accuracy, and further examine component reweighting and recurrent layer sharing. Our main empirical observations are:

- Our proposed component reweighting (§4.2) can improve performance while changing the balance between bias and diversity (§5.4).
- Models that reuse layers exhibit substantially lower measured layer diversity, suggesting that diversity provides a useful perspective for analyzing recurrent layer architectures (§5.6).

Overall, our results suggest that the benefit of additional Transformer components depends not only on the information contributed by each component individually, but also on how complementary their contributions are. This provides a layer-level perspective on one aspect of the diminishing returns observed as model scale increases.

2 Residual Stream in Transformers

As shown in Figure 1, in pre-layer normalization type transformers (Xiong et al., 2020) commonly used in LLMs, there is a residual stream (Elhage et al., 2021) consisting of embedding layers, multi-head attention (MHA), and multi-layer perceptron (MLP). In this residual stream, the outputs of each module are added together, and finally, the normalized result is projected to predict the output. Letting $L = (\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(|L|)})$, $E = (\mathbf{e}^{(1)}, \cdots, \mathbf{e}^{(|E|)})$, $M = (\mathbf{m}^{(1)}, \cdots, \mathbf{m}^{(|M|)})$, and $A = (\boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(|A|)})$ be sequences of entire layers, embedding layers, MLP layers, and MHA layers, respectively, the logit prediction on the residual stream is represented as follows:

$$\mathbf{logits} = W_{out}\mathrm{LN}(\sum_{i=1}^{|L|} \boldsymbol{z}^{(i)}) = W_{out}\mathrm{LN}(\sum_{j=1}^{|E|} \boldsymbol{e}^{(j)} + \sum_{k=1}^{|M|} \boldsymbol{m}^{(k)} + \sum_{l=1}^{|A|} \boldsymbol{a}^{(l)}), \tag{1}$$

where LN is RMSNorm (Zhang & Sennrich, 2019) and W_{out} is a projection layer for vocabulary V. By letting the scaling by RMSNorm be $\mathbf{s}$ and $W_{out}(\mathbf{s} \odot \mathbf{z}^{(i)})$ be $\mathbf{u}^{(i)}$, we can reformulate Equation 1 as follows (Chang et al., 2024):

$$(1) \equiv \mathbf{logits} = W_{out}(\mathbf{s} \odot \sum_{i=1}^{|L|} \mathbf{z}^{(i)}) = \sum_{i=1}^{|L|} W_{out}(\mathbf{s} \odot \mathbf{z}^{(i)}) = \sum_{i=1}^{|L|} \mathbf{u}^{(i)}. \tag{2}$$

We use this equation to investigate the theoretical effects of modules in the residual stream.

3 Theoretical Analysis

In this section, we analyze the theoretical characteristics of the residual stream of Transformer layers.

3.1 Prediction Discrepancy

In general, Transformers output the token with the highest prediction probability for a given input. This can be formalized as follows by letting SM be a softmax function and $\bar{\mathbf{u}} = \frac{1}{|L|}\sum_{i=1}^{|L|}\mathbf{u}^{(i)}$:

$$\arg\max_y SM(\mathbf{logits}) = \arg\max_y \mathbf{logits} = \arg\max_y \frac{1}{|L|}\sum_{i=1}^{|L|}\mathbf{u}^{(i)} = \arg\max_y \bar{\mathbf{u}}. \tag{3}$$

By defining $\mathbf{u}^*$ as the true distribution of logits $\bar{\mathbf{u}}$, we can consider a prediction error between Transformer layers and its true distribution on logits as the following Mean Squared Error (MSE):

$$MSE(\mathbf{u}^*, \bar{\mathbf{u}}) = \frac{1}{|V|}\sum_{j=1}^{|V|}(u_j^* - \bar{u}_j)^2 = \mathbb{E}_{j\in V}[(u_j^* - \bar{u}_j)^2]. \tag{4}$$

We use this error to analyze the discrepancy between predicted and true logits and reveal theoretical insights into the residual streams of Transformer layers.

3.1.1 Importance of Diversity

First, we focus on the diversity arising from the differences between the predictions of each Transformer layer. From Eqs. 1, 2, and 4, we can decompose the discrepancy between the Transformer's predictions and the true logit distribution into bias and diversity terms as in the following theorem.

Theorem 1 (**Bias and Diversity Decomposition**). *The prediction error of Transformer layers, $MSE(\hat{\boldsymbol{u}}, \bar{\boldsymbol{u}})$, can be decomposed into bias and diversity (ambiguity) (Wood et al., 2024) terms (Krogh & Vedelsby, 1994) as follows (See Appendix C.1 for the proof.):*

$$(4) = \underbrace{\mathbb{E}_{j\in V}[\mathbb{E}_{\mathbf{u}\in L}[(u_j^* - u_j)^2]]}_{\textit{Bias}} - \underbrace{\mathbb{E}_{j\in V}[\mathbb{E}_{\mathbf{u}\in L}[(\bar{u}_j - u_j)^2]]}_{\textit{Diversity}} \tag{5}$$

$$= \frac{1}{|L|}\mathbb{E}_{j\in V}[\underbrace{|E|\mathbb{E}_{\mathbf{e}\in E}[(u_j^* - u_j^{\boldsymbol{e}})^2]}_{\textit{Embedding Bias}} + |M|\underbrace{\mathbb{E}_{\mathbf{m}\in M}[(u_j^* - u_j^{\boldsymbol{m}})^2]}_{\textit{MLP Bias}} + |A|\underbrace{\mathbb{E}_{\boldsymbol{a}\in A}[(u_j^* - u_j^{\boldsymbol{a}})^2]}_{\textit{Attention Bias}}] \tag{6}$$

$$- \frac{1}{|L|}\mathbb{E}_{j\in V}[|E|\underbrace{\mathbb{E}_{\mathbf{e}\in E}[(\bar{u}_j - u_j^{\boldsymbol{e}})^2]}_{\textit{Embedding Diversity}} + |M|\underbrace{\mathbb{E}_{\mathbf{m}\in M}[(\bar{u}_j - u_j^{\boldsymbol{m}})^2]}_{\textit{MLP Diversity}} + |A|\underbrace{\mathbb{E}_{\boldsymbol{a}\in A}[(\bar{u}_j - u_j^{\boldsymbol{a}})^2]}_{\textit{Attention Diversity}}], \tag{7}$$

where $\mathbf{u}^{\boldsymbol{e}}$, $\mathbf{u}^{\boldsymbol{a}}$, and $\mathbf{u}^{\boldsymbol{m}}$ are $\mathbf{u}$ at the same position of $\boldsymbol{e}$, $\boldsymbol{a}$, and $\boldsymbol{m}$, respectively.

As shown in Eq. 5, MSE can be decomposed into two terms: bias and diversity. Since the diversity term is negative, we can see that when diversity increases, the discrepancy from the true distribution decreases, meaning that prediction accuracy improves. This is different from the bias-variance decomposition, which is widely known in the field of machine learning.

Equations 6 and 7 show the further decomposition of the bias and diversity terms corresponding to the modules on the residual stream in Eq. 1. From this, we can see that each module contributes to both bias and diversity, and that the number of modules weighs the contribution. Therefore, we can confirm that MLP layers and attention layers play an important role in Transformers in terms of the number of layers. The contribution of the embedding layer is small based on the number of layers; however, its relative importance may increase in lightweight models. We empirically investigate these hypotheses through experiments in §5.2.

3.1.2 IMPORTANCE OF MODULES

Up to this point, our analysis has assumed that the outputs of individual modules are added uniformly to the residual stream. In practice, however, the vectors produced by different modules have varying magnitudes, and thus their contributions to the residual stream are not uniform. Motivated by this observation, we focus on the norms of each module's output vector, leading to the following theorem.

Theorem 2 (**Weighted Voting**). *Letting a unit vector for* $\mathbf{u}^{(i)}$ *be* $\hat{\mathbf{u}}^{(i)} = \mathbf{u}^{(i)}/\|\mathbf{u}^{(i)}\|$ *and its importance on a residual stream be* $\alpha_i = \|\mathbf{u}^{(i)}\|/\sum_{i'=1}^{|L|}\|\mathbf{u}^{(i')}\|$*, we can reformulate Eq. 3 as* $\arg\max \mathbf{logits} = \arg\max \sum_{i=1}^{|L|} \alpha_i \hat{\mathbf{u}}^{(i)}$*. (See Appendix C.2 for the proof.)*

Theorem 2 shows that summing the outputs of each module in a residual stream can be interpreted as weighted voting. Furthermore, we can apply this weighting to the bias and diversity decomposition as in the following theorem.

Theorem 3 (**Weighted Bias and Diversity**). *Based on Theorem 2, we can reformulate Bias and Diversity terms in Eq. 7 as follows (Krogh & Vedelsby, 1994)*[1] *(See Appendix C.3 for the proof.):*

$$\underbrace{\mathbb{E}_{j\in V}[\mathbb{E}_{(\hat{\mathbf{u}},\alpha)}[\alpha(u_j^*-u_j)^2]]}_{Bias} - \underbrace{\mathbb{E}_{j\in V}[\mathbb{E}_{(\hat{\mathbf{u}},\alpha)}[\alpha(\bar{u}_j-u_j)^2]]}_{Diversity} \tag{8}$$

Theorem 3 shows that although we can adjust the Bias and Diversity terms by appropriately choosing α, it is still difficult to improve one without affecting the other due to the shared weight. This adjustment corresponds to the component reweighting method Chang et al. (2024), which estimates the importance of each module using training data. We discuss the relationship between Theorems 2 and 3 and component reweighting in §4.2.

3.1.3 LIMITATION OF DIVERSITY

As discussed in §3.1.1, the diversity term is important for improving prediction performance. However, there are limitations to its effectiveness, as shown in the next theorem.

Theorem 4 (**Limitation of Diversity**). *For the aggregate decomposition in Eq. 5,*

$$0 \leq Diversity \leq Bias. \tag{9}$$

Consequently, $Bias \to 0$ *implies* $Diversity \to 0$*. (See Appendix C.4 for the proof.)*

Theorem 4 formalizes a simple constraint: disagreement cannot remain large when the average component error vanishes. It does not imply the converse, nor does it state that high component error automatically produces useful diversity. Note that the two quantities can be additionally coupled through shared second-order terms. (See Appendix B.1 for the details.)

3.2 BEYOND LOGIT-SPACE DECOMPOSITION: INFORMATION-THEORETIC DIVERSITY

The bias-diversity analysis above exploits the additive structure of residual-stream contributions in logit space. However, the notion of complementary information among layers is not inherently restricted to logits. We therefore introduce an information-theoretic formulation that treats each layer output as a random variable and characterizes the information jointly provided by these variables about the target. This formulation does not require linear aggregation and can be instantiated using either logit-space representations or their corresponding softmax probability distributions. Although we retain the notation $\mathbf{u}^{(i)}$ below for continuity with the preceding analysis, the information-theoretic results depend only on the induced random variables rather than on the additive logit-space decomposition as shown in the following theorem:

Theorem 5 (**Generalized diversity**). *Letting* Y *be a true label, variables* $\mathbf{u}^{(1)}, \cdots, \mathbf{u}^{(|L|)}$ *be* $U_{1:|L|}$*, and* $g(U_{1:|L|})$ *be a function*[2] *predicting the best label. Its prediction error* $p(Y \neq g(U_{1:|L|}))$ *satisfies*

[1] Note that $\bar{\mathbf{u}} = \sum_{i=1}^{|L|} \alpha_i \hat{\mathbf{u}}^{(i)}$ in this formulation.

[2] This function must follow the Bayes decision rule Hellman & Raviv (1970). In the case of the Transformer, the Bayesian decision rule stands only when the output distribution of the Transformer fits that of its training data, and argmax is used for its prediction. Since we target scaling laws that require large parameters with large data, this premise is valid and natural. Note that even a one-layer Transformer can satisfy the Bayesian decision rule (Shen et al., 2025; Nguyen & Nguyen-Tang, 2025).

the following inequality (Brown, 2009; Zhou & Li, 2010) (See Appendix C.6 for the proof.):

$$\frac{H(Y) - I(U_{1:|L|}; Y) - 1}{\log |Y|} \leq p(Y \neq g(U_{1:|L|})) \leq \frac{H(Y) - I(U_{1:|L|}; Y)}{2}, \tag{10}$$

where $H(Y)$ is entropy and $I(U_{1:|L|}; Y)$ is joint mutual information. To minimize the error, we should maximize $I(U_{1:|L|}; Y)$ which can be decomposed to (Zhou & Li, 2010):

$$I(U_{1:|L|}; Y) = \underbrace{\sum_{i=1}^{|L|} I(\mathbf{u}^{(i)}; Y)}_{Relevancy} + \underbrace{\underbrace{\mathcal{I}(U_{1:|L|}|Y)}_{Conditional\ Redundancy} - \underbrace{\mathcal{I}(U_{1:|L|})}_{Redundancy}}_{Information\ Theoretic\ Diversity}, \tag{11}$$

where $\mathcal{I}(U_{1:|L|})$ is total correlation and $\mathcal{I}(U_{1:|L|}|Y)$ is conditional total correlation.

Theorem 5 provides an information-theoretic analogue of the bias-diversity view: *Relevancy* captures information that individual layer outputs provide about the target, while *Information-Theoretic Diversity* characterizes their dependence structure. We use this decomposition to analyze how prediction-error bounds change as additional layer outputs are included.

3.2.1 Stacking Layers

Increasing depth is one way to increase model capacity, but the number of layers is not equivalent to the total parameter count. We therefore do not attempt to derive parameter scaling laws directly. Instead, we ask whether the information contributed by additional layer outputs exhibits diminishing marginal returns that are qualitatively consistent with scaling behavior. We analyze this question through the following information-theoretic results.

Theorem 6 (Monotonicity of Bounds). *In Eq. 11, when $U_{1:|L|}$ increases (that means new layers are added to $U_{1:|L|}$), the lower bound $\frac{H(Y)-I(U_{1:|L|};Y)-1}{\log |Y|}$, and the upper bound $\frac{H(Y)-I(U_{1:|L|};Y)}{2}$ of the prediction error are non-increasing (See Appendix C.7 for the proof.)*

Theorem 6 shows that, within this formulation, adding an additional layer output cannot worsen the corresponding information-theoretic prediction-error bounds.[3] The next theorem characterizes a regime in which the marginal changes diminish.

Theorem 7 (Diminishing Returns). *Let $\mathcal{G} = \{1, \ldots, |L|\}$ index the layer outputs, let $U_S = \{\mathbf{u}^{(j)} : j \in S\}$ for $S \subseteq \mathcal{G}$, and define the set function $F(S) = I(U_S; Y)$. For any $S \subseteq T \subseteq \mathcal{G}$ and $i \notin T$, let $C = T \setminus S$ and*

$$\epsilon_{i,S,T} = \Big[I(\mathbf{u}^{(i)}; U_C \mid Y, U_S) - I(\mathbf{u}^{(i)}; U_C \mid U_S) \Big]_+, \tag{12}$$

where $[x]_+ = \max\{x, 0\}$. Then

$$F(S \cup \{i\}) - F(S) \geq F(T \cup \{i\}) - F(T) - \epsilon_{i,S,T}. \tag{13}$$

Thus, if $\epsilon_{i,S,T} \leq \epsilon$ for all S, T, i, then F is ϵ-approximately submodular. Moreover, the lower and upper prediction-error bounds in Eq. 10, denoted by $B_{\text{low}}(S)$ and $B_{\text{up}}(S)$, satisfy

$$B_{\text{low}}(S \cup \{i\}) - B_{\text{low}}(S) \leq B_{\text{low}}(T \cup \{i\}) - B_{\text{low}}(T) + \frac{\epsilon}{\log |Y|}, \tag{14}$$

$$B_{\text{up}}(S \cup \{i\}) - B_{\text{up}}(S) \leq B_{\text{up}}(T \cup \{i\}) - B_{\text{up}}(T) + \frac{\epsilon}{2}. \tag{15}$$

In particular, if $I(\mathbf{u}^{(i)}; U_C \mid U_S) \geq I(\mathbf{u}^{(i)}; U_C \mid Y, U_S)$ for all S, T, i, then $\epsilon = 0$ and the exact submodular/supermodular diminishing-return relations hold. (See Appendix C.8 for the proof.)

Theorem 7 quantifies departures from diminishing returns through $\epsilon_{i,S,T}$, which measures the excess dependence induced by conditioning on the target. When this term is small, marginal target information and the corresponding prediction-error bounds exhibit approximate diminishing returns; when it is zero, the relation is exact.[4] This set-function result compares the same candidate component under nested conditioning sets and therefore does not by itself imply decreasing gains along the natural layer order; we examine only qualitative consistency with model depth in §5.5.

[3] This statement concerns the bounds for an expanding set of layer outputs and does not by itself guarantee higher empirical accuracy for a separately trained deeper model.

[4] See Appendix B.2 for another case.

4 REMAINING PROBLEMS

4.1 DIFFERENCE BETWEEN PROBABILITY AND LOGITS

Previous discussions have focused on the logits in the output layer of the Transformer. However, when actually using the Transformer, the output is selected from vectors normalized by the softmax layer. Therefore, to investigate whether our theoretical findings are practical, it is desirable to conduct experiments using actual models and datasets. To investigate the behavior of MSE and bias through experiments, we should prepare the true logit distribution, $\hat{\mathbf{u}}$. However, assuming the correct answers in softmax, softmax returns the same value for shifts in the input like $\frac{e^{\mathbf{logits}_i}}{\sum_j e^{\mathbf{logits}_j}} = \frac{e^{(\mathbf{logits}_i - \mu)}}{\sum_j e^{(\mathbf{logits}_j - \mu)}}$, resulting in an infinite number of true logits, which is difficult to handle. To address this issue, we approximate MSE and bias in Eq. 5 using the following equation:

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) \approx MSE(\tilde{\mathbf{u}}, softmax(\bar{\mathbf{u}})); bias \approx \frac{1}{|V|}\sum_{i=1}^{|V|}\frac{1}{|L|}\sum_{j=1}^{|L|}\left(\widetilde{u}_i - softmax(\mathbf{u}^{(j)})_i\right)^2, \quad (16)$$

where $\tilde{\mathbf{u}}$ represents a one-hot vector whose dimension of a gold label is one and others are zero.

4.2 COMPONENT REWEIGHTING AND VECTOR NORMS

Conflict Between Two Types of Importance The original component reweighting Chang et al. (2024) predicts a label y from $\mathbf{x} = \{\mathbf{u}^{(1)}, \cdots, \mathbf{u}^{(|L|)}\}$ by adjusting the importance of modules using learnable parameters $\boldsymbol{w}$ trained through the negative log-likelihood over the training dataset D_{train}:

$$P(y|\mathbf{x}) = softmax\left(\frac{\sum_{i=1}^{|L|} w_i u_y^{(i)}}{\sum_{j=1}^{|V|}\sum_{i=1}^{|L|} w_i u_j^{(i)}}\right), \quad (17) \qquad \sum_{(y,\mathbf{x})\in D_{\text{train}}} -\log P(y \mid \mathbf{x}). \quad (18)$$

By focusing on the norms of the input vectors, as in Theorem 2, we can reformulate Eq. 17 as

$$P(y|\mathbf{x}) = softmax\left(\frac{\sum_{i=1}^{|L|} w_i \alpha_i \hat{u}_y^{(i)}}{\sum_{j=1}^{|V|}\sum_{i=1}^{|L|} w_i \alpha_i \hat{u}_j^{(i)}}\right). \quad (19)$$

where α_i denotes the norm of the output vector from the i-th module and $\hat{u}^{(i)}$ is its normalized version. This reformulation reveals that the original component reweighting conflates two different notions of importance: (1) vector norms $\boldsymbol{\alpha}$ derived from pretrained modules and (2) learnable weights $\boldsymbol{w}$ estimated from task-specific data.

Solution To mitigate this issue, we propose two solutions. The first one is normalizing the output vectors of each module to remove the influence of vector norms from the prediction as in Equation 20. However, this solution entirely discards the prior importance information encoded in the pretrained modules. The second one can handle this problem by incorporating the vector norms as a prior via maximum a posteriori (MAP) estimation as in Equation 21:

$$P(y|\mathbf{x}) = softmax\left(\frac{\sum_{i=1}^{|L|} w_i \hat{u}_y^{(i)}}{\sum_{j=1}^{|V|}\sum_{i=1}^{|L|} w_i \hat{u}_j^{(i)}}\right), \quad (20) \qquad \sum_{(y,\mathbf{x})\in D_{train}} -\log P(y|\mathbf{x}) + \lambda\|\boldsymbol{w} - \boldsymbol{\alpha}\|, \quad (21)$$

where λ is a hyperparameter controlling the strength of the regularization term. We investigate the effectiveness of these solutions in §5.4.

4.3 OUR ASSUMPTION FOR THE PARAMETER SCALING LAWS

Dimension Size Our previous discussion on the parameter scaling laws has been based on the assumption that the number of layers in a Transformer is proportional to the number of parameters. However, when considering the relationship with the parameter scaling laws in a rigorous manner, it is necessary to take into account elements at the neuron level corresponding to the dimension size. Fortunately, it is known that the residual stream in Transformers can be decomposed down to the neuron level (Elhage et al., 2021), allowing us to replace "layers" with "neurons" in our discussion and maintain the validity of our claims.

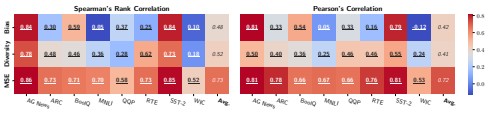

Figure 2: The correlation between accuracy and MSE, bias, and diversity of models across tasks. The underlined scores show the statistical significance ($p < 0.05$).[5] Note that the scores on Avg. are not the target of the significance test.

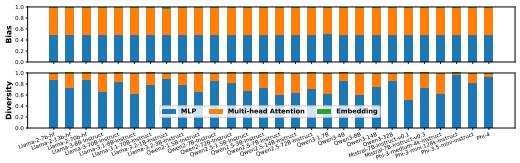

Figure 3: The average proportion of bias and diversity accounted for by each module in each model in all datasets.

Reuse of Layers Another oversight is the existence of models that recursively utilize layers. In such cases, the number of layers does not necessarily correspond to the number of parameters. A well-known example is the Universal Transformer (Dehghani et al., 2019), whose efficiency is also recognized in natural language generation tasks (Takase & Kiyono, 2023). Recently, there has been a trend to repurpose such models as small-scale LLMs (Touvron et al., 2023). We verify whether the increase in layers independent of parameters in such models is consistent with our claim in §5.6.

5 Empirical Analysis

General Settings We used the following Transformer models from HuggingFace Transformers (Wolf et al., 2020): MobileLLM (Liu et al., 2024); Llama-2 (Touvron et al., 2023); Llama-3, Llama-3.1, Llama-3.2, and Llama-3.3 Grattafiori et al. (2024); Phi-3 and Phi-3.5 Abdin et al. (2024a); Phi-4 (Abdin et al., 2024b); Mistral (Jiang et al., 2023); Qwen2 Yang et al. (2024), Qwen 2.5 Qwen et al. (2025), and Qwen 3 Yang et al. (2025). By following the settings of the previous work (Chang et al., 2024), we used at most the first 2000 instances in the AG News (Zhang et al., 2015), ARC (Clark et al., 2018), BoolQ (Clark et al., 2019), MNLI (Williams et al., 2018), QQP (Wang et al., 2017), RTE, SST-2, and WIC (Wang et al., 2018) datasets. We used accuracy as a metric, predefined templates (Bach et al., 2022) for prompting and restricting the output (See Appendix D.1 for the details).

5.1 Correlation of MSE, Bias, and Diversity with Performance

To clarify the relationship between MSE, bias, diversity, and performance on real tasks, we computed the Pearson and Spearman correlations between accuracy and MSE, bias, and diversity at each junction point in the residual stream. Figure 2 presents the results.[6] As shown in the figure, MSE exhibits a strong correlation with accuracy. In contrast, bias and diversity show moderate correlations, with trends that vary across tasks, suggesting that the model changes the roles of individual layers depending on the task. The higher Spearman correlation than the Pearson correlation for MSE indicates that MSE better captures prediction outcomes (i.e., correctness) than prediction confidence.

5.2 Influence of each Module

Figure 3 shows the influence of each module on bias and diversity based on the weights corresponding to their layer size, represented by Eqs. 6 and 7. In bias, the contributions for each module are proportional to their layer size. This is the reason for the few contributions of the embedding layers. These tendencies are almost the same across the models. In contrast, the MLP layers dominate diversity. However, the tendency differs depending on the model. This result suggests the necessity of improving the diversity of multi-head attention layers. It supports the previous work, indicating the existence of unnecessary attention heads in Transformers (He et al., 2024). Similar to the case of the bias, the contribution of embedding layers is limited.

[5] This is based on Student's t-test (Student, 1908).

[6] Unlike accuracy and diversity, lower bias and MSE indicate better performance. To reflect this, we used the negative values of bias and MSE when computing the correlations. In addition, to align the value ranges, we standardized each score for each model. Using the z-transformation Corey et al. (1998), we report the averaged correlations (Avg.).

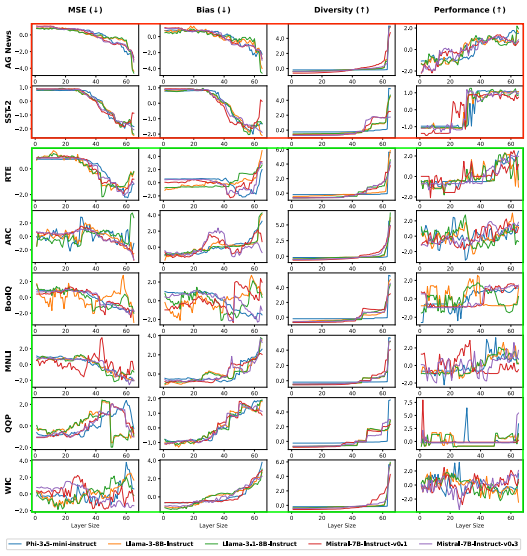

Figure 4: The relationship between MSE, bias, diversity, and performance on each task. Note that each metric is standardized to capture the change in terms of the number of layers.

Table 1: Reweighting results in accuracy. Bold indicates the best score for each model. † denotes the statistical significance Koehn (2004) (p¡0.05).

		AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.
Meta-Llama-3-8B-Instruct	None	35.2	27.7	62.8	34.9	65.4	47.3	85.0	50.0	51.0
	Base	60.8	38.8	75.2	**52.1**	65.5	71.5	93.3	**55.3**	64.1
	Map	**64.2**†	37.4	**77.7**†	50.4	71.2†	69.0	93.0	50.0	64.1
	Unit	62.0	**39.1**	75.3	51.3	**72.6**†	**74.0**	**93.6**	54.5	**65.3**†
Llama-3.1-8B-Instruct	None	30.2	27.7	62.8	41.1	65.4	47.3	88.6	50.0	51.6
	Base	61.9	**32.6**	83.6	**50.6**	65.7	70.0	90.1	56.3	63.9
	Map	64.4†	28.8	**83.6**	47.5	**75.2**†	65.7	89.4	58.2	64.1
	Unit	**65.8**†	31.9	83.5	50.2	75.1†	**72.9**	**90.5**	**59.9**	**66.2**†
Qwen2-7B-Instruct	None	30.1	72.5	62.8	26.4	65.4	46.9	50.9	50.2	50.7
	Base	**65.2**	**83.7**	83.7	**66.9**	79.7	81.6	93.3	63.5	**77.2**
	Map	59.4	81.2	83.5	65.0	**81.4**†	82.3	**93.6**	63.5	76.2
	Unit	59.3	80.7	**83.7**	66.3	80.6	**83.4**†	93.2	**65.5**	76.6
Qwen2.5-7B-Instruct	None	29.6	81.4	62.8	30.7	65.4	42.2	55.3	50.3	52.2
	Base	**67.8**	**88.8**	82.6	70.8	83.9	**86.3**	93.9	**68.8**	**80.4**
	Map	65.2	86.1	82.2	**74.3**†	84.0	82.3	93.7	67.1	79.4
	Unit	66.1	86.3	**83.0**	74.2†	**84.1**	85.2	**94.2**	67.7	80.1
Qwen2.5-72B-Instruct	None	37.8	43.2	64.1	44.0	65.4	47.3	50.9	50.0	50.3
	Base	65.4	**96.0**	90.9	76.2	85.3	**86.6**	93.8	73.4	83.4
	Map	64.2	95.4	90.9	76.3	85.1	85.6	94.3	**73.8**	83.2
	Unit	**66.7**	95.8	**91.0**	**80.2**†	**85.4**	86.3	**94.7**†	73.2	**84.2**†
Qwen3-14B	None	51.2	20.0	65.5	36.2	65.4	47.3	81.2	50.0	52.1
	Base	63.2	84.6	87.8	59.5	**83.7**	84.1	93.7	60.5	77.1
	Map	63.4	**85.1**	88.1	63.2†	82.6	83.8	93.6	60.8	77.6
	Unit	**68.4**†	84.9	**88.2**	**66.1**†	83.1	**85.2**	**94.3**	**63.8**†	**79.3**†
Qwen3-32B	None	45.4	32.3	69.8	29.8	62.2	47.3	70.4	49.8	50.9
	Base	58.0	87.5	83.4	69.0	85.7	**83.4**	91.6	70.2	78.6
	Map	60.6†	**92.5**†	83.5	71.1†	85.7	81.2	**92.5**	70.5	79.7†
	Unit	**65.9**†	92.3†	**84.4**	**72.9**†	**85.9**	81.9	92.5	**71.0**	**80.8**†
Mistral-7B-Instruct-v0.1	None	48.6	33.7	62.7	37.4	65.4	49.5	68.3	51.3	52.1
	Base	62.6	39.6	72.4	40.0	65.4	**75.8**	89.2	52.7	62.2
	Map	63.2	36.1	75.1†	43.3†	65.5†	72.9	88.8	54.5	62.4
	Unit	**67.4**†	**40.4**	**75.4**†	**44.9**†	**66.5**†	74.0	**89.4**	**57.5**†	**64.4**†
Mistral-7B-Instruct-v0.3	None	51.3	32.3	63.6	42.1	55.4	56.0	78.4	53.3	54.0
	Base	65.5	**43.0**	80.0	52.4	65.4	77.6	89.7	59.2	66.6
	Map	65.7	40.7	80.7	50.7	**66.4**†	**79.1**	89.7	61.3	66.8
	Unit	**69.7**†	42.5	**81.3**†	**54.2**†	66.0†	78.7	**91.5**†	**62.2**	**68.3**†

5.3 Trade-off Relationship between Bias and Diversity

We investigate whether there is actually a trade-off between bias and diversity, as shown in Theorems 4 and 8. Figure 4 shows the results.[7] As these results show, when performance does not improve with additional layers, there is a trade-off between bias and diversity (red frame), whereas when performance improves, this trade-off no longer holds (green frame). This is consistent with the fact that bias and diversity share only one term, as shown in Theorem 8. The limitation of diversity shown in Theorem 4 is caused by bias being close to zero, but in reality, bias rarely approaches zero. Therefore, the decrease in bias does not hinder the improvement of diversity. Therefore, although a trade-off relationship may exist in some cases, it is possible to improve both bias and diversity.

5.4 Module Reweighting

To investigate the effectiveness of component reweighting approaches and clarify their relationship to the underlying bias and diversity, we compare the methods in Eqs. 17 and 18 (Base), our methods in Eq. 20 (Unit) and Eq. 21 (Map), and the setting without module reweighting (None). (See Appendix D.2 for details.)

Table 1 reports the accuracies for the top-10 models based on the original accuracy.[9] From the performance improvements over None, we can see the effectiveness of module reweighting. Interestingly, in most cases (light green cells), our Map and Unit, outperform the baseline, Base, whereas Base sometimes outperforms

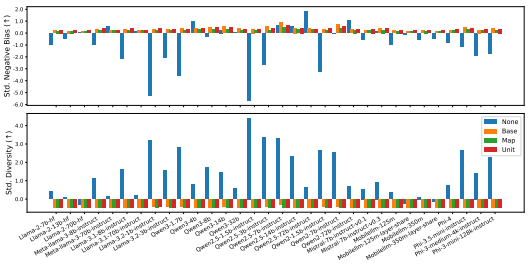

Figure 5: Changes in bias and diversity through module reweighting. We report z-score–normalized values, and biases are negated.[8]

[7] Note that due to low visibility, this figure only covers the top-1 model and four models with the same number of layers that achieved the top five accuracy on average across all tasks. See Appendix E.1 for the overall result.

[8] To avoid the differences in value ranges caused by reweighting, we adopt softmax in the diversity calculation.

[9] The results of all models are reported in Appendix E.2.

Table 2: The results of MobileLLM. The suffix -LS indicates the model recurrently shares its layers.

MobileLLM	Accuracy									MSE ($\times 10^2$)								
	AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.	AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.
125M	49.4	27.2	62.6	30.4	43.8	50.9	53.1	53.0	46.3	18.5	18.7	24.9	22.3	25.0	25.0	25.0	25.0	23.0
125M-LS	47.9	28.4	62.7	30.4	34.8	52.7	55.4	52.7	45.6	18.7	18.7	25.0	22.2	25.0	25.0	25.0	25.0	23.1
350M	50.4	27.7	62.6	30.8	35.2	53.4	52.6	55.2	46.0	18.5	18.7	24.9	22.2	25.0	25.0	25.0	25.0	23.0
350M-LS	52.4	25.6	59.7	30.4	34.6	52.7	49.1	50.0	44.3	18.7	18.8	25.0	22.2	25.3	25.0	25.0	25.0	23.1
MobileLLM	Bias ($\times 10^2$)									Diversity ($\times 10^2$)								
	AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.	AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.
125M	18.6	18.8	25.1	22.4	25.7	25.2	25.0	25.3	23.3	6.2	10.1	11.8	8.1	20.3	6.2	21.8	13.1	12.2
125M-LS	18.7	18.8	25.0	22.3	25.2	25.1	25.0	25.0	23.1	1.3	1.6	3.0	0.8	2.3	1.5	2.6	1.1	1.8
350M	18.6	18.8	25.0	22.3	25.5	25.1	25.0	25.2	23.2	9.6	9.2	16.7	8.0	3.9	10.4	54.0	12.1	15.5
350M-LS	18.7	18.8	25.1	22.3	25.5	25.1	25.1	25.1	23.2	2.9	4.6	8.6	3.2	2.6	3.0	5.7	3.4	4.2

Map and Unit (light red cells). These results indicate a collision between the estimated weights and the norms of module vectors in LLMs. This observation is further supported by Unit, which, not considering module vector norms, outperforms Map. Also, some cases in which Base outperforms Map and Unit highlight the importance of using the module vector norms as prior information.

To better understand these performance changes, we investigate the bias and diversity terms underlying the predictions. Figure 5 shows the results. The figure clearly demonstrates that component reweighting converts diversity gains into bias improvements. This trade-off relationship is consistent with Theorem 3, which indicates the difficulty of improving only one of the bias and diversity terms.

5.5 Cross-Family Depth Trends

Figure 6 compares average task accuracy with the reported number of residual components for several public model families (model list in Appendix D.3). Within these selected families, larger checkpoints often show higher average accuracy and smaller incremental gains at the high end. This pattern is qualitatively consistent with the diminishing marginal information gain in Theorem 9.

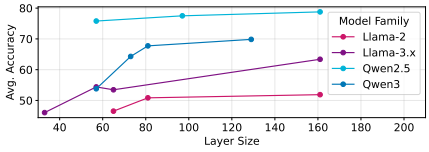

Figure 6: Performance and Layer count.

5.6 Models that Reuse Layers

We examine models that reuse layers using MobileLLM. Table 2 shows that recurrent layer-sharing variants exhibit substantially lower measured diversity and, in these settings, somewhat lower average accuracy. This association is compatible with our diversity-based perspective, but it does not establish that reduced diversity causes the performance difference. The results motivate further controlled study of whether preserving complementary layer contributions can improve recurrent layer architectures.

6 Conclusion

In this paper, we analyzed Transformer residual-stream components through bias-diversity decomposition and characterized how component accuracy, diversity, and importance jointly affect the aggregated prediction. We further developed an information-theoretic view in which adding layer outputs cannot worsen prediction-error bounds and derived an approximate diminishing-returns result for dependent layer outputs. The departure from exact submodularity is controlled by how much conditioning on the target increases dependence between a candidate component and additional observed components. This provides a layer-level mechanism that is qualitatively consistent with one aspect of parameter scaling laws, diminishing returns, rather than a derivation of the scaling law itself. Across multiple tasks and public LLM families, our empirical results show patterns compatible with this perspective, including systematic relationships among bias, diversity, and accuracy, changes in the bias-diversity balance under component reweighting, and reduced measured diversity under recurrent layer sharing. Overall, layer diversity provides a useful diagnostic lens for studying how Transformer components contribute as model depth increases.

Ethics Statement

This work studies the internal behavior of pretrained Transformer language models through theoretical analysis and experiments on existing NLP benchmarks. We do not collect new data, conduct experiments involving human participants, or introduce a new dataset. Our experiments use publicly available pretrained models and established benchmark datasets, and our analysis is intended to improve scientific understanding of how Transformer components contribute to model predictions. Our results concern prediction behavior, component diversity, and model architecture, and should not be interpreted as establishing properties related to fairness, social bias, privacy, safety, or reliability in deployment. In particular, the datasets and pretrained models considered in this work may inherit biases or other limitations from their original data sources and training procedures; we do not attempt to characterize or mitigate these issues here. The component-analysis and reweighting techniques studied in this paper could potentially be applied to models used in downstream applications, so their use in high-stakes settings would require application-specific evaluation of safety, fairness, privacy, and other relevant risks. To the best of our knowledge, the present work does not introduce additional ethical concerns beyond those associated with the pretrained models and benchmark datasets on which it is evaluated.

AI Use Statement

Generative AI tools were used to assist with manuscript review, language editing, and drafting portions of the paper, including the Ethics Statement and AI Use Statement. Generative AI was also used to review Theorem 7 and its proof, including checking the mathematical reasoning and identifying possible issues or points requiring clarification. The authors independently verified the theorem, proof, and all AI-generated suggestions before incorporating them into the manuscript. Generative AI was not used to generate experimental data or to replace the authors' own evaluation and interpretation of the experimental results. The authors take full responsibility for the correctness and content of the paper, including all theoretical claims, proofs, experimental results, and text.

Reproducibility Statement

Experimental settings are described in §5 and Appendix D. These sections provide the information required to reproduce our results, including templates, models, decoding methods, and GPU environment details. We will release our code to support reproducibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou.

Phi-3 technical report: A highly capable language model locally on your phone, 2024a. URL https://arxiv.org/abs/2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024b. URL https://arxiv.org/abs/2412.08905.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. PromptSource: An integrated development environment and repository for natural language prompts. In Valerio Basile, Zornitsa Kozareva, and Sanja Stajner (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 93–104, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-demo.9. URL https://aclanthology.org/2022.acl-demo.9/.

Gavin Brown. An information theoretic perspective on multiple classifier systems. In *Proceedings of the 8th International Workshop on Multiple Classifier Systems*, MCS '09, pp. 344–353, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 9783642023255. doi: 10.1007/978-3-642-02326-2_35. URL https://doi.org/10.1007/978-3-642-02326-2_35.

Gavin Brown, Jeremy L. Wyatt, and Peter TiŇo. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6(55):1621–1650, 2005. URL http://jmlr.org/papers/v6/brown05a.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. When parts are greater than sums: Individual LLM components can outperform full models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10280–10299, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.574. URL https://aclanthology.org/2024.emnlp-main.574/.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

David M Corey, William P Dunlap, and Michael J Burke. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *The Journal of general psychology*, 125(3):245–261, 1998.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Satoru Fujishige. Polymatroidal dependence structure of a set of random variables. *Information and Control*, 39(1):55–72, 1978. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(78)91063-X. URL https://www.sciencedirect.com/science/article/pii/S001999587891063X.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya

Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. What matters in transformers? not all attention is needed, 2024. URL https://arxiv.org/abs/2406.15786.

Martin Hellman and Josef Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL https://arxiv.org/abs/1712.00409.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu (eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250/.

Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, pp. 324–331, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/b8c37e33defde51cf91e1e03e51657da-Paper.pdf.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. Mobilellm: optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Quan Nguyen and Thanh Nguyen-Tang. One-layer transformers are provably optimal for in-context reasoning and distributional association learning in next-token prediction tasks, 2025. URL https://arxiv.org/abs/2505.15009.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of gaussian mixtures, 2025. URL https://arxiv.org/abs/2410.11778.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2331554.

Milan Studeny. *Probabilistic conditional independence structures*. Springer Science & Business Media, 2006.

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Orpf8yDjdj.

Sho Takase and Shun Kiyono. Lessons on parameter sharing across layers in transformers. In Nafise Sadat Moosavi, Iryna Gurevych, Yufang Hou, Gyuwan Kim, Young Jin Kim, Tal Schuster, and Ameeta Agrawal (eds.), *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 78–90, Toronto, Canada (Hybrid), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.sustainlp-1.5. URL https://aclanthology.org/2023.sustainlp-1.5/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 4144–4150, 2017. doi: 10.24963/ijcai.2017/579. URL https://doi.org/10.24963/ijcai.2017/579.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language

processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *J. Mach. Learn. Res.*, 24(1), mar 2024. ISSN 1532-4435.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Zhi-Hua Zhou and Nan Li. Multi-information ensemble diversity. In Neamat El Gayar, Josef Kittler, and Fabio Roli (eds.), *Multiple Classifier Systems*, pp. 134–144, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-12127-2.

A Limitations

A.1 Nested Representations versus Retrained Models

Theorems 6 and 9 assume a fixed joint distribution in which an existing representation set is enlarged by adding another variable. Training a deeper or wider Transformer generally changes all representations and the data/optimization trajectory. The theory therefore does not imply monotonic improvement for independently trained checkpoints and should not be interpreted as a derivation of parameter scaling laws.

A.2 Conditional-Independence Assumption

The diminishing-returns result in Theorem 9 assumes that component representations are conditionally independent given the target label. Successive Transformer components share the same input and are computationally dependent, so this assumption may be strongly violated in practice. We use it to obtain a transparent limiting-case result; directly estimating conditional dependence among high-dimensional component representations is left for future work.

A.3 Logit-Space Theory versus Probability-Space Proxy

The exact ambiguity identity is stated for vectors in a common logit space and a reference target vector. Supervised datasets do not define a unique target logit vector, so the experiments use the probability-space proxy in Eq. 16. Because softmax is nonlinear, empirical agreement with the proxy is qualitative support rather than a direct numerical verification of the logit-space theorem.

A.4 Lack of a Controlled Depth-Scaling Family

The comparison in Figure 6 is not a controlled depth ablation. Checkpoints differ in width, parameter count, data, optimization, and sometimes model generation; Llama-3.x, for example, combines Llama-3.2 and Llama-3.1 checkpoints. We therefore do not use Figure 6 to validate Theorem 9 or to claim a causal relationship between depth and performance. Its role is only to show that the public-checkpoint trend is qualitatively consistent with the theoretical prediction.

Ideally, the effect of depth would be tested using models trained with matched width, data, optimization, and compute allocation while varying only depth. The MobileLLM layer-sharing comparison partially decouples repeated computation from parameter count, but it is not a substitute for such a controlled depth-scaling experiment.

A.5 Limited to Inference Time

Our theoretical and empirical analyses target internal behavior during inference. Training dynamics (Sun et al., 2026), including how diversity emerges during optimization, are outside the scope of this work.

A.6 Autoregressive Error Accumulation

The analysis focuses on a single prediction step. Error accumulation and changing residual-stream statistics across long autoregressive generations are not modeled.

A.7 Proprietary Models

Our empirical analyses require access to internal hidden states and are therefore restricted to open models. The theoretical algebra is architecture-level, but its empirical assumptions have not been tested on proprietary systems.

A.8 Computational Resources

The breadth of model families limits the extent of controlled retraining experiments. We prioritize transparent analysis of publicly available checkpoints and identify controlled depth scaling as the primary experiment needed to strengthen the causal claim.

B Additional Theorems

B.1 Bias and Diversity Trade-off

Theorem 8 (**Bias and Diversity Trade-off**). *Bias and Diversity in the decomposition of the prediction error in Transformer layers (Eq. 5) depend on each other as follows (Brown et al., 2005):*

$$Bias = \overline{bias}^2 + \Omega; Diversity = \Omega - \left[\frac{1}{|L|}\overline{var} + \left(1 - \frac{1}{|L|}\right)\overline{cov}\right], \tag{22}$$

where $\overline{bias} = \mathbb{E}_{\mathbf{u}\in L}[\mathbb{E}_{j\in V}[u_j] - \mathbb{E}_{j\in V}[u_j^*]], \overline{var} = \mathbb{E}_{\mathbf{u}\in L}[\mathbb{E}_{j\in V}(u_j - \mathbb{E}_{k\in V}[u_k])^2], \Omega = \overline{var} + \mathbb{E}_{\mathbf{u}\in L}[(\mathbb{E}_{j\in V}[u_j] - \mathbb{E}_{j\in V}[\overline{u}_j])^2], \overline{cov} = \frac{2}{|L|(|L|-1)}\sum_{1\le i<j\le |L|}\mathrm{Cov}_{k\in V}(u_k^i, u_k^j)$*, and* Cov *is covariance. (See Appendix C.5 for the proof.)*

By focusing on Eq. 22, we can see that the bias and diversity terms share Ω. This suggests that attempting to reduce the bias term may also reduce Ω, which in turn may reduce the diversity term. In other words, Theorem 8 demonstrates the general trade-off relationship between the bias and diversity term, emphasizing the difficulty of maximizing the diversity term without affecting the bias term.

B.2 Convergence

The next theorem also characterizes a regime in which the marginal changes diminish.

Theorem 9 (**Convergence**). *When* $\mathbf{u}^{(i)} \in U_{1:|L|}$ *are independent given* Y*,* $I(U_{1:|L|}; Y)$ *is submodular and non-decreasing on* $U_{1:|L|}$ *and the lower bound* $\frac{H(Y)-I(U_{1:|L|};Y)-1}{\log|Y|}$ *and the upper bound* $\frac{H(Y)-I(U_{1:|L|};Y)}{2}$ *are supermodular and non-increasing. (See Appendix C.9 for the proof.)*

Under the conditional-independence assumption in Theorem 9, the submodular and supermodular behavior implies diminishing marginal changes in the information quantity and the corresponding prediction-error bounds as layer outputs are added. This behavior is qualitatively consistent with the diminishing returns observed in parameter scaling laws. However, the theorem does not derive a parameter scaling law or imply that collaboration between layers necessarily becomes more difficult with depth; rather, it identifies one theoretical regime in which diminishing returns arise.

C Proofs

C.1 The proof for Theorem 1

According to the previous work Krogh & Vedelsby (1994), when the coefficients of each term in Eq. 4 satisfy the definition of probability, we can decompose Eq. 4 into Eq. 5. Here, the coefficients follow a uniform distribution, and the decomposition of Eq. 5 holds. We can induce Eqs. 6 and 7 based on the fact that the weighted mean equals the original mean.

C.2 The proof for Theorem 2

We can prove Theorem 2 by reformulating Eq. 3 as follows:

$$(3) = \arg\max_y \sum_{i=1}^{|L|} \mathbf{u}^{(i)} = \arg\max_y \sum_{i=1}^{|L|} \|\mathbf{u}^{(i)}\| \frac{\mathbf{u}^{(i)}}{\|\mathbf{u}^{(i)}\|} = \arg\max_y \sum_{i=1}^{|L|} \|\mathbf{u}^{(i)}\| \hat{\mathbf{u}}^{(i)}$$
$$= \arg\max_y \sum_{i=1}^{|L|} \frac{\|\mathbf{u}^{(i)}\|}{\sum_{i=1}^{|L|} \|\mathbf{u}^{(i)}\|} \hat{\mathbf{u}}^{(i)} = \arg\max \sum_{i=1}^{|L|} \alpha_i \hat{\mathbf{u}}^{(i)}. \tag{23}$$

C.3 The proof for Theorem 3

The previous work Krogh & Vedelsby (1994) shows the decomposition of MSE for weighted ensembling, which requires the following conditions: (1) the sum of all weights equals 1, and (2) each weight is non-negative. In our case, α_i is based on vector norms and is therefore always non-negative. Moreover, the definition

$$\alpha_i = \frac{\|\mathbf{u}^{(i)}\|}{\sum_{i=1}^{|L|} \|\mathbf{u}^{(i)}\|}, \tag{24}$$

guarantees that the weights sum to 1. Hence, Theorem 3 is proved.

C.4 The proof for Theorem 4

Both MSE and Diversity in Eq. 5 are averages of squared quantities and are therefore non-negative. Rearranging Eq. 5 gives

$$Diversity = Bias - MSE. \tag{25}$$

Hence $0 \leq Diversity \leq Bias$. If $Bias \to 0$, the squeeze theorem gives $Diversity \to 0$.

C.5 The proof for Theorem 8

The decomposition of Krogh & Vedelsby (1994) is a sufficient condition for the decomposition of Brown et al. (2005). Since we proved that Theorem 4 holds, Theorem 8 also holds.

C.6 The proof for Theorem 5

See Zhou & Li (2010) for the proof of the bound and decomposition. Since their claim does not restrict the elements of the input and output in their formulation, we can apply their bound and decomposition to the behavior of the residual stream in Transformers.

C.7 The proof for Theorem 6

First, we prove the following lemma:

Lemma 10 (**Monotonicity of each Term**). *In Eq. 11, when $U_{1:|L|}$ increases (that means new layers are added to $U_{1:|L|}$), Relevancy, Conditional Redundancy, and Redundancy monotonically increase whereas Information Theoretic Diversity does not monotonically increase or decrease.*

Since *Relevancy* is a sum of joint mutual information. We can consider its increase when $U_{1:|L|}$ obtains a new element, $\mathbf{u}^{(|L|+1)}$ as follows:

$$\sum_{i=1}^{|L|+1} I(\mathbf{u}^{(i)}; Y) - \sum_{i=1}^{|L|} I(\mathbf{u}^{(i)}; Y) \tag{26}$$
$$= I(\mathbf{u}^{(|L|+1)}; Y). \tag{27}$$

Since $I(\mathbf{u}^{(|L|+1)}; Y)$ is non-negative, *Relevancy* monotonically increases.

Conditional Redundancy is conditional total correlation (conditional multi-information). We can consider its increase when $U_{1:|L|}$ obtains a new element, $\mathbf{u}^{(|L|+1)}$ as follows:

$$\mathcal{I}(U_{1:|L|+1}|Y) - \mathcal{I}(U_{1:|L|}|Y) \tag{28}$$

$$= \left(\sum_{i=1}^{|L|+1} H(\mathbf{u}^{(i)}|Y) \right) - H(U_{1:|L|+1}|Y) - \left(\sum_{i=1}^{|L|} H(\mathbf{u}^{(i)}|Y) \right) + H(U_{1:|L|}|Y) \tag{29}$$

$$= H(\mathbf{u}^{(|L|+1)}|Y) - H(U_{1:|L|+1}|Y) + H(U_{1:|L|}|Y) \tag{30}$$

$$= H(\mathbf{u}^{(|L|+1)}|Y) - \Big(H(\mathbf{u}^{(|L|+1)}|Y) + H(U_{1:|L|}|Y, \mathbf{u}^{(|L|+1)}) \Big) + H(U_{1:|L|}|Y) \tag{31}$$

$$= H(U_{1:|L|}|Y) - H(U_{1:|L|}|Y, \mathbf{u}^{(|L|+1)}) \tag{32}$$

$$= I(U_{1:|L|}; \mathbf{u}^{(|L|+1)}|Y) \tag{33}$$

$$= I(U_{1:|L|+1}|Y). \tag{34}$$

Since $I(U_{1:|L|+1}|Y)$ is non-negative, *Conditional Redundancy* monotonically increases.

Redundancy is total correlation (multi-information). We can similarly consider its increase when $U_{1:|L|}$ obtains a new element, $\mathbf{u}^{(|L|+1)}$ as follows:

$$\mathcal{I}(U_{1:|L|+1}) - \mathcal{I}(U_{1:|L|}) \tag{35}$$

$$= \left(\sum_{i=1}^{|L|+1} H(\mathbf{u}^{(i)}) \right) - H(U_{1:|L|+1}) - \left(\sum_{i=1}^{|L|} H(\mathbf{u}^{(i)}) \right) + H(U_{1:|L|}) \tag{36}$$

$$= H(\mathbf{u}^{(|L|+1)}) - H(U_{1:|L|+1}) + H(U_{1:|L|}) \tag{37}$$

$$= H(\mathbf{u}^{(|L|+1)}) - \Big(H(\mathbf{u}^{(|L|+1)}) + H(\mathbf{u}^{(|L|+1)}|U_{1:|L|}) \Big) + H(U_{1:|L|}) \tag{38}$$

$$= H(U_{1:|L|}) - H(\mathbf{u}^{(|L|+1)}|U_{1:|L|}) \tag{39}$$

$$= I(U_{1:|L|+1}). \tag{40}$$

Since $I(U_{1:|L|+1})$ is non-negative, *Redundancy* monotonically increases. Note that the previous work Studeny (2006) introduces the monotonicity of total correlation.

Regarding *Information Theoretic Diversity*, we can consider its increase when $U_{1:|L|}$ obtains a new element, $\mathbf{u}^{(|L|+1)}$ by utilizing Eqs. 34 and 40 as follows:

$$\big(\mathcal{I}(U_{1:|L|+1}|Y) - \mathcal{I}(U_{1:|L|}|Y) \big) - \big(\mathcal{I}(U_{1:|L|+1}) - \mathcal{I}(U_{1:|L|}) \big) \tag{41}$$

$$= I(U_{1:|L|+1}|Y) - I(U_{1:|L|+1}). \tag{42}$$

When elements in $U_{1:|L|+1}$ are independent but not independent given Y, Eq. 42 is non-negative. On the other hand, when elements in $U_{1:|L|+1}$ are not independent but are independent given Y, Eq. 42 is non-positive. Thus, *Information Theoretic Diversity* does not have monotonicity. Hence, Lemma 10 is proved.

Finally, using Eqs. 27 and 42, we can similarly consider the increase of $I(U_{1:|L|}; Y)$ when $U_{1:|L|}$ obtains a new element, $\mathbf{u}^{(|L|+1)}$ as follows

$$I(U_{1:|L|+1}; Y) - I(U_{1:|L|}; Y) \tag{43}$$

$$= I(\mathbf{u}^{(|L|+1)}; Y) + I(U_{1:|L|+1}|Y) - I(U_{1:|L|+1}) \tag{44}$$

$$= I(\mathbf{u}^{(|L|+1)}; Y|U_{1:|L|}). \tag{45}$$

Here, $I(\mathbf{u}^{(|L|+1)}; Y|U_{1:|L|})$ is non-negative. Next, we prove the monotonicity of the bounds. Since $H(Y)$ and $\log |Y|$ are constant with the increase of $U_{1:|L|}$ and $I(U_{1:|L|}; Y)$ satisfy the monotonicity by Eq. 45, the lower bound $\frac{H(Y) - I(U_{1:|L|}; Y) - 1}{\log |Y|}$ and the upper bound $\frac{H(Y) - I(U_{1:|L|}; Y)}{2}$ of the prediction error monotonically decreases. Therefore, Theorem 6 is proved.

Therefore, Theorem 6 is proved.

C.8 THE PROOF FOR THEOREM 7

Let $F(S) = I(U_S; Y)$ and define the marginal target information of component i given S by

$$\Delta_i(S) = F(S \cup \{i\}) - F(S) = I(Y; \mathbf{u}^{(i)} \mid U_S). \tag{46}$$

The final equality follows from the chain rule of mutual information. Because conditional mutual information is non-negative, $\Delta_i(S) \geq 0$, so F is non-decreasing without any independence assumption.

For $S \subseteq T$, $i \notin T$, and $C = T \setminus S$, we have $U_T = (U_S, U_C)$. The difference between the two marginal contributions is

$$\begin{aligned}
\Delta_i(S) - \Delta_i(T) &= I(Y; \mathbf{u}^{(i)} \mid U_S) - I(Y; \mathbf{u}^{(i)} \mid U_S, U_C) && (47)\\
&= H(\mathbf{u}^{(i)} \mid U_S) - H(\mathbf{u}^{(i)} \mid Y, U_S) \\
&\quad - H(\mathbf{u}^{(i)} \mid U_S, U_C) + H(\mathbf{u}^{(i)} \mid Y, U_S, U_C) && (48)\\
&= I(\mathbf{u}^{(i)}; U_C \mid U_S) - I(\mathbf{u}^{(i)}; U_C \mid Y, U_S). && (49)
\end{aligned}$$

Equation 49 is an exact identity and does not assume independence among layer outputs. By the definition of $\epsilon_{i,S,T}$ in Eq. 12,

$$\Delta_i(S) - \Delta_i(T) \geq -\epsilon_{i,S,T}, \tag{50}$$

which is equivalent to Eq. 13. If $\epsilon_{i,S,T} \leq \epsilon$ uniformly, F is therefore ϵ-approximately submodular. If

$$I(\mathbf{u}^{(i)}; U_C \mid U_S) \geq I(\mathbf{u}^{(i)}; U_C \mid Y, U_S), \tag{51}$$

then $\epsilon_{i,S,T} = 0$ and exact submodularity follows.

It remains to show the corresponding statement for the prediction-error bounds. Define

$$B_{\text{low}}(S) = \frac{H(Y) - F(S) - 1}{\log |Y|}, \qquad B_{\text{up}}(S) = \frac{H(Y) - F(S)}{2}. \tag{52}$$

Since F is non-decreasing, both bounds are non-increasing. Furthermore,

$$B_{\text{low}}(S \cup \{i\}) - B_{\text{low}}(S) = -\frac{\Delta_i(S)}{\log |Y|}, \tag{53}$$

$$B_{\text{up}}(S \cup \{i\}) - B_{\text{up}}(S) = -\frac{\Delta_i(S)}{2}. \tag{54}$$

Applying Eq. 13 to these expressions yields Eqs. 14 and 15. Hence, the prediction-error bounds are approximately supermodular with slack $\epsilon / \log |Y|$ and $\epsilon/2$, respectively, completing the proof of Theorem 7.

C.9 THE PROOF FOR THEOREM 9

We can decompose Theorem 9 into the following two lemmas:

Lemma 11 (**Submodularity of each Term**). *In Eq. 11, when the variables in $U_{1:|L|}$ are independent given Y, $I(U_{1:|L|}; Y)$ and Information Theoretic Diversity are submodular, and $I(U_{1:|L|}; Y)$ is non-decreasing on $U_{1:|L|}$. (See Appendix C.9.1 for the proof.)*

Lemma 12 (**Supermodularity of Upper and Lower Bounds**). *In Eq. 10, when the variables in $U_{1:|L|}$ are independent given Y, the lower bound $\frac{H(Y) - I(U_{1:|L|}; Y) - 1}{\log |Y|}$ and the upper bound $\frac{H(Y) - I(U_{1:|L|}; Y)}{2}$ are supermodular and non-increasing on $U_{1:|L|}$. (See Appendix C.9.2 for the proof.)*

Based on the following proofs regarding these lemmas, Theorem 9 is proved.

C.9.1 THE PROOF FOR LEMMA 11

Because $I(U_{1:|L|}; Y)$ is a joint mutual information, it is submodular and non-decreasing on $U_{1:|L|}$ under the conditional dependence of the variables in $U_{1:|L|}$ given Y (Krause & Guestrin, 2005). To

Table 3: Dataset names and licenses.

Dataset Name	License
AG News	CC BY 4.0
ARC	CC BY-SA 4.0
BoolQ	CC BY-SA 3.0
MNLI	CC BY 4.0
QQP	Apache 2.0
RTE	Unknown
SST-2	Unknown
WiC	CC BY-NC-SA 3.0

check that, we define the change of $I(U_{1:|L|};Y)$ when adding $\mathbf{u}$ to $U_{1:|L|}$ as follows:

$$\begin{aligned}\Delta(\mathbf{u}|U_{1:|L|}) =& I(U_{1:|L|},\mathbf{u}^{(|L|+1)};Y) - I(U_{1:|L|};Y) && (55)\\ =& I(Y,\mathbf{u}|U_{1:|L|}) + I(U_{1:|L|};Y) - I(U_{1:|L|};Y) && (56)\\ =& I(Y,\mathbf{u}|U_{1:|L|}) && (57)\end{aligned}$$

Since $I(Y,\mathbf{u}|U_{1:|L|})$ is no less than zero, $I(U_{1:|L|};Y)$ monotonically increases when $U_{1:|L|}$ increases under the conditional dependence of the variables in $U_{1:|L|}$ given Y. By utilizing this reformulation,

$$I(Y,\mathbf{u}|U_{1:|L|}) = H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|Y,U_{1:|L|}), \tag{58}$$

and utilizing entropy's submodularity, we can induce the following inequality:

$$\begin{aligned}&\Delta(\mathbf{u}|U_{1:|L|}) - \Delta(\mathbf{u}|U_{1:|L|+1}) && (59)\\ =& H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|Y,U_{1:|L|}) - \big(H(\mathbf{u}|U_{1:|L|+1}) - H(\mathbf{u}|Y,U_{1:|L|+1})\big) && (60)\\ =& H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|U_{1:|L|+1}) - \big(H(\mathbf{u}|Y,U_{1:|L|}) - H(\mathbf{u}|Y,U_{1:|L|+1})\big) && (61)\\ =& H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|U_{1:|L|+1}) - (H(\mathbf{u}|Y) - H(\mathbf{u}|Y)) && (62)\\ =& H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|U_{1:|L|+1}) && (63)\\ \geq& 0 && (64)\end{aligned}$$

Therefore, $I(U_{1:|L|};Y)$ is submodular on $U_{1:|L|}$ under the conditional dependence of the variables in $U_{1:|L|}$ given Y.

Information Theoretic Divergence is a sum of total correlation and conditional total correlation. Both functions are supermodular according to the previous works Fujishige (1978); Studeny (2006). Since $\mathcal{I}(U_{1:|L|}|Y)$ is zero under the given condition, *Information Theoretic diversity* becomes

$$-\mathcal{I}(U_{1:|L|}). \tag{65}$$

Here $-\mathcal{I}(U_{1:|L|})$ is submodular and thus, *Information Theoretic diversity* is a submodular function on $U_{1:|L|}$.

C.9.2 The proof for Lemma 12

Under the condition, since $I(U_{1:|L|};Y)$ is a submodular function, $-I(U_{1:|L|};Y)$ is a supermodular function. Only $-I(U_{1:|L|};Y)$ is a term including $U_{1:|L|}$ in both bounds. Therefore, the upper and lower bounds are supermodular functions and monotonically decrease.

D Experimental Details

D.1 Overall Settings

Table 3 shows the datasets and their licenses. Excluding the AG New dataset, we used a validation split because the test set has not been released. Table 4 shows the templates used for each task. All

Table 4: The used templates for each task.

Task	Template	Options
AG News	What label best describes this news article?\n {text} {answer}	World, Sports, Business, Science
ARC	Pick the most correct option to answer the following question.\n\n{text} \n\n{options} {answer}	A B C D
BoolQ	{passage}\n\nAfter reading this passage, I have a question: {question} True or False? {answer}	False, True
MNLI	Suppose it's true that {text1} Then, is "{text2}" {options} {answer}	Always, Sometimes, Never
QQP	I'm an administrator on the website Quora. There are two posts, one that asks "{question1}" and another that asks "{question2}" I can merge questions if they are asking the same thing. Can I merge these two questions? {answer}	no, yes
RTE	{text1} Using only the above description and what you know about the world, is {text2} definitely correct? Yes or No? {answer}	Yes, No
SST-2	{text}\nQuestion: Was that sentence positive or negative? Answer: {answer}	negative, positive
WIC	Does the word "{word}" have the same meaning in these two sentences? Yes, No?\n {sentence1}\n{sentence2} {answer}	No, Yes

templates are extracted from promptsource (Bach et al., 2022)[10]. The output vocabulary is restricted to the tokens used in options corresponding to the task. Table 5 shows the details of the used models in our experiments. In the implementation, we modified the released code from Chang et al. (2024)[11]. We used an NVIDIA RTX A6000, which has 48GB of VRAM, to run the models. We applied 4-bit quantization to 70B models and 8-bit quantization to the 32B model. Each setting can be completed in under 4 hours using a single NVIDIA RTX A6000.

D.2 Component Reweighting

We adopted the experimental settings reported by Chang et al. (2024). Accordingly, we used SGD with a learning rate of 5×10^{-2} and included a regularization term. Since our preliminary experiments showed that L2 regularization is more effective than L1 regularization, we used L2 for regularization. We set both the regularization coefficient and λ to 1×10^{-2}. Furthermore, following their test-set usage, we selected the first 2,000 instances from each training dataset to estimate the weights. Training was conducted for a maximum of 2,000 epochs with a patience of 3. In the statistical significance test, we used paired-bootstrap resampling Koehn (2004) with 10000 trials, using the original dataset size.

D.3 Performance Change by Increasing Layers

The models used in §5.5 are as follows:

- Llama-2: Llama-2-7b-hf, Llama-2-13b-hf, Llama-2-70b-hf

[10] https://github.com/bigscience-workshop/promptsource
[11] https://github.com/terarachang/LLMDecomp

Table 5: The statistics of the used models.

Model Name	Parameter Size	Layer Size	License
meta-llama/Llama-3.2-1B-Instruct	1B	33	LLaMA 3.2 License
Qwen/Qwen2-1.5B-Instruct	2B	57	Apache 2.0
Qwen/Qwen2-7B-Instruct	8B	57	Apache 2.0
Qwen/Qwen2.5-1.5B-Instruct	2B	57	Apache 2.0
Qwen/Qwen2.5-7B-Instruct	8B	57	Apache 2.0
Qwen/Qwen3-1.7B	2B	57	Apache 2.0
meta-llama/Llama-3.2-3B-Instruct	3B	57	LLaMA 3.2 License
microsoft/Phi-3-mini-128k-instruct	3.8B	65	MIT
microsoft/Phi-3.5-mini-instruct	3.8B	65	MIT
meta-llama/Llama-2-7b-hf	7B	65	LLaMA 2 License
mistralai/Mistral-7B-Instruct-v0.1	7B	65	Apache 2.0
mistralai/Mistral-7B-Instruct-v0.3	7B	65	Apache 2.0
meta-llama/Meta-Llama-3-8B-Instruct	8B	65	LLaMA 3 License
meta-llama/Llama-3.1-8B-Instruct	8B	65	LLaMA 3.1 License
Qwen/Qwen2.5-3B-Instruct	3B	73	Qwen Research License
Qwen/Qwen3-4B	4B	73	Apache 2.0
Qwen/Qwen3-8B	8B	73	Apache 2.0
Qwen/Qwen3-14B	14.8B	81	Apache 2.0
meta-llama/Llama-2-13b-hf	13B	81	LLaMA 2 License
microsoft/Phi-3-medium-4k-instruct	14B	81	MIT
microsoft/phi-4	14.7B	81	MIT
Qwen/Qwen2.5-14B-Instruct	14.7B	97	Apache 2.0
Qwen/Qwen3-32B	32.8B	129	Apache 2.0
Qwen/Qwen2-72B-Instruct	73B	161	Tongyi Qianwen License
Qwen/Qwen2.5-72B-Instruct	73B	161	Qwen License
meta-llama/Llama-2-70b-hf	70B	161	LLaMA 2 License
meta-llama/Llama-3.1-70B-Instruct	71B	161	LLaMA 3.1 License
meta-llama/Meta-Llama-3-70B-Instruct	71B	161	LLaMA 3 License
facebook/MobileLLM-125M	125M	61	MIT
facebook/MobileLLM-125M-layer-share	125M	121	MIT
facebook/MobileLLM-350M	350M	65	MIT
facebook/MobileLLM-350M-layer-share	350M	129	MIT

- Llama-3.x: Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct
- Qwen2.5: Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-72B-Instruct
- Qwen3: Qwen/Qwen3-1.7B, Qwen/Qwen3-8B, Qwen/Qwen3-14B, Qwen/Qwen3-32B

We selected model families that include at least three models, after excluding small language models with more layers and fewer parameters than the models in the same family.

E Detailed Results

E.1 Trade-off Relationship

Figure 7 shows the detailed results of the MSE, bias, diversity, and performance of LLMs for each layer.

E.2 Reweighting Layers

Table 6 shows the detailed results of the module reweighting.

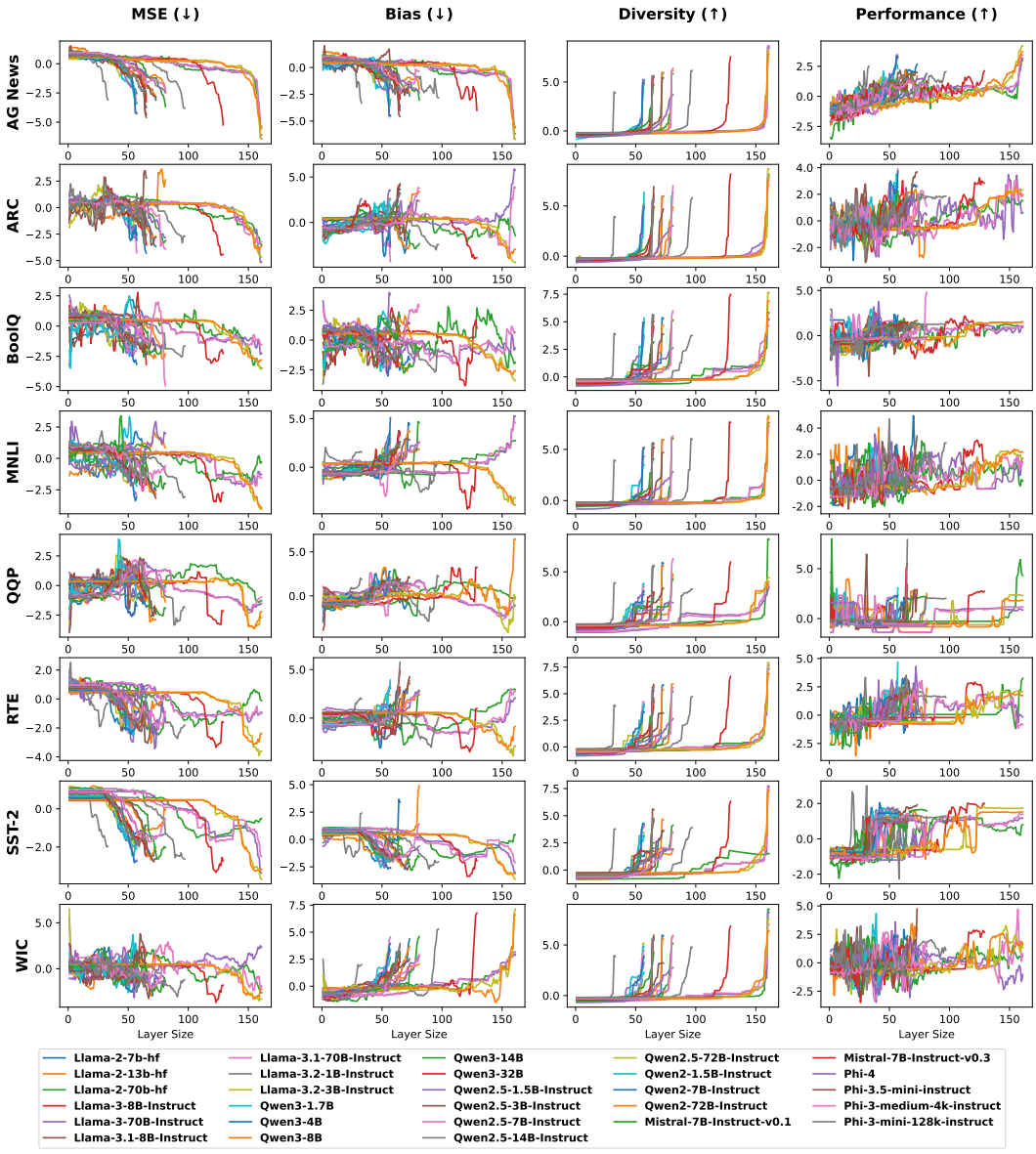

Figure 7: The relationship between MSE, bias, diversity, and performance of all models on each task. The settings are the same as Figure 4.

Table 6: Module reweighting results in accuracy. The notations are the same as in Table 1.

		AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.
Qwen2.5-1.5B-Instruct	None	35.9	34.9	62.5	33.2	65.4	53.1	50.9	50.0	48.2
	Base	64.1	**61.8**	**73.7**	**55.6**	65.2	77.3	89.7	**61.6**	**68.6**
	Map	66.5†	60.9	70.3	50.8	67.3	77.3	**90.3**	55.3	67.3
	Unit	**67.1**†	61.4	71.6	52.2	**68.8**	**78.3**	90.0	58.0	68.4
Qwen2.5-3B-Instruct	None	28.0	27.4	62.8	36.5	65.4	47.3	50.9	50.0	46.0
	Base	62.2	**67.5**	**75.2**	60.5	79.6	80.9	**92.1**	**61.3**	72.4
	Map	64.2†	65.6	73.0	60.2	**80.3**	80.9	91.7	58.2	71.8
	Unit	**68.6**†	67.2	73.2	**62.7**†	78.9	**81.2**	91.5	60.5	**73.0**
Llama-2-7b-hf	None	54.4	**28.9**	38.2	29.6	65.4	46.2	71.4	52.7	48.4
	Base	61.8	28.8	62.8	34.5	65.4	57.4	59.7	54.9	53.2
	Map	62.7	28.1	62.4	35.2	66.1†	**61.4**	**83.7**†	**59.6**†	**57.4**†
	Unit	**63.3**	26.8	**62.9**	**36.7**†	**68.3**†	57.8	82.9†	58.3	57.1†
Llama-2-13b-hf	None	44.1	21.8	39.9	35.1	65.4	47.3	50.9	49.8	44.3
	Base	70.1	26.1	62.8	34.9	65.4	46.6	81.4	54.5	55.2
	Map	70.0	25.3	**63.1**	41.0†	65.3	54.2	85.8†	**57.7**	57.8†
	Unit	**71.9**†	**30.5**†	63.1	**42.8**†	**66.5**†	**58.1**	**88.3**†	57.1	**59.8**†
Llama-2-70b-hf	None	46.7	28.6	39.6	30.6	35.6	47.3	52.1	47.6	41.0
	Base	62.7	40.9	62.8	37.7	65.4	62.5	83.1	51.3	58.3
	Map	68.7†	47.5†	67.4†	47.0†	74.9†	68.6†	89.4†	62.4†	65.7†
	Unit	**71.1**†	**50.0**†	**68.0**†	**47.9**†	**77.5**†	**69.3**†	**90.9**†	**63.0**†	**67.2**†
Meta-Llama-3-8B-Instruct	None	35.2	27.7	62.8	34.9	65.4	47.3	85.0	50.0	51.0
	Base	60.8	38.8	75.2	**52.1**	65.5	71.5	93.3	**55.3**	64.1
	Map	**64.2**†	37.4	**77.7**†	50.4	71.2†	69.0	93.0	50.0	64.1
	Unit	62.0	**39.1**	75.3	51.3	**72.6**†	**74.0**	**93.6**	54.5	**65.3**†
Meta-Llama-3-70B-Instruct	None	31.0	27.4	37.2	34.8	22.5	39.0	50.2	56.3	37.3
	Base	66.1	38.9	84.5	58.7	83.1	**83.4**	93.0	55.2	70.4
	Map	70.9†	47.0†	86.2†	61.3†	82.5	78.3	94.2†	61.1†	72.7†
	Unit	**71.9**†	**49.8**†	**86.7**†	**63.5**†	**84.6**†	79.4	**94.7**†	**64.6**†	**74.4**†
Llama-3.1-8B-Instruct	None	30.2	27.7	62.8	41.1	65.4	47.3	88.6	50.0	51.6
	Base	61.9	**32.6**	83.6	**50.6**	65.7	70.0	90.1	56.3	63.9
	Map	64.4†	28.8	**83.6**	47.5	**75.2**†	65.7	89.4	58.2	64.1
	Unit	**65.8**†	31.9	83.5	50.2	75.1†	**72.9**	**90.5**	**59.9**	**66.2**†
Llama-3.1-70B-Instruct	None	32.3	24.4	37.4	34.7	19.1	37.9	75.5	51.1	39.0
	Base	62.9	56.0	86.0	55.6	84.2	82.7	**93.3**	66.3	73.4
	Map	69.9†	**59.5**†	**87.7**†	60.5†	84.2	79.8	92.4	65.2	74.9
	Unit	**72.0**†	58.9	87.6†	**63.2**†	**85.0**†	**83.0**	92.3	**67.2**	**76.2**†
Llama-3.2-1B-Instruct	None	40.8	22.8	62.6	36.4	**65.4**	50.2	50.9	50.3	47.4
	Base	**66.4**	25.4	**62.8**	40.5	65.4	**68.6**	78.8	**56.9**	**58.1**
	Map	61.3	27.2	62.1	39.3	65.4	51.6	**79.7**	56.7	55.4
	Unit	61.9	**27.9**	62.8	**42.3**†	65.4	54.2	72.4	56.4	55.4
Llama-3.2-3B-Instruct	None	47.6	28.1	66.1	35.3	34.7	59.9	51.0	46.6	46.2
	Base	65.0	**31.8**	**79.4**	50.1	65.4	**66.8**	88.0	51.3	62.2
	Map	65.0	25.4	79.2	49.5	**70.4**†	66.1	88.0	54.5	62.3
	Unit	**65.9**	31.8	77.4	**52.2**†	65.4	65.0	**89.3**	**56.0**†	**62.9**
Qwen2-1.5B-Instruct	None	26.3	32.3	62.4	34.9	65.4	47.3	50.9	50.0	46.2
	Base	69.2	**48.4**	**69.3**	**50.2**	65.4	**74.7**	88.9	**58.2**	**65.5**
	Map	67.0	**45.1**	67.0	47.2	**65.7**	68.2	90.3	57.4	63.5
	Unit	**70.1**	46.5	67.9	49.8	65.0	70.8	**90.5**†	58.2	64.8
Qwen2-7B-Instruct	None	30.1	72.5	62.8	26.4	65.4	46.9	50.9	50.2	50.7
	Base	**65.2**	**83.7**	83.7	**66.9**	79.7	81.6	93.3	63.5	**77.2**
	Map	59.4	81.2	83.5	65.0	**81.4**†	82.3	**93.6**	63.5	76.2
	Unit	59.3	80.7	**83.7**	66.3	80.6	**83.4**†	93.2	**65.5**	76.6
Qwen2-72B-Instruct	None	28.1	26.7	60.0	37.0	61.0	47.3	76.8	50.0	48.4
	Base	66.6	87.5	89.0	74.1	82.7	82.7	93.8	**72.6**	81.1
	Map	64.6	90.4†	89.2	75.9†	82.8	83.8	94.0	71.3	81.5
	Unit	**67.2**	**91.2**†	**89.2**	**76.2**†	**84.2**†	**84.1**	**94.5**	71.9	**82.3**†
Qwen2.5-7B-Instruct	None	29.6	81.4	62.8	30.7	65.4	42.2	55.3	50.3	52.2
	Base	**67.8**	**88.8**	82.6	70.8	83.9	**86.3**	93.9	**68.8**	**80.4**
	Map	65.2	86.1	82.2	**74.3**†	84.0	82.3	93.7	67.1	79.4
	Unit	66.1	86.3	**83.0**	74.2†	**84.1**	85.2	**94.2**	67.7	80.1
Qwen2.5-14B-Instruct	None	26.2	36.3	54.8	27.6	65.4	47.3	50.9	50.2	44.8
	Base	**67.3**	92.6	**88.1**	71.2	85.0	86.6	**93.5**	**69.4**	81.7
	Map	64.5	93.3	87.1	73.0†	84.5	86.3	92.7	68.2	81.2
	Unit	65.5	**94.2**†	88.0	**74.7**†	**85.9**†	**88.1**	93.5	68.2	**82.3**

		AG News	ARC	BoolQ	MNLI	QQP	RTE	SST-2	WIC	Avg.
Qwen2.5-72B-Instruct	None	37.8	43.2	64.1	44.0	65.4	47.3	50.9	50.0	50.3
	Base	65.4	**96.0**	90.9	76.2	85.3	**86.6**	93.8	73.4	83.4
	Map	64.2	95.4	90.9	76.3	85.1	85.6	94.3	**73.8**	83.2
	Unit	**66.7**	95.8	**91.0**	**80.2**†	**85.4**	86.3	**94.7**†	73.2	**84.2**†
Qwen3-1.7B	None	43.2	28.1	63.3	32.9	65.4	47.3	51.0	50.0	47.6
	Base	65.8	**47.2**	71.4	41.7	65.4	**76.9**	88.8	**58.0**	64.4
	Map	69.1†	47.2	**74.8**†	50.2†	**66.9**†	72.2	88.8	54.2	65.4
	Unit	**70.3**†	46.1	74.7†	**51.5**†	65.5†	71.1	**89.8**	56.0	**65.6**
Qwen3-4B	None	43.0	25.8	63.6	44.2	65.0	47.3	51.1	50.0	48.8
	Base	64.9	50.2	76.1	54.9	81.4	75.1	90.5	56.4	68.7
	Map	68.3†	**57.7**†	80.4†	**61.0**†	81.1	75.5	90.8	**58.6**	71.7†
	Unit	**70.0**†	57.7†	**81.3**†	60.8†	**81.6**	**77.6**	**90.9**	58.6	**72.3**†
Qwen3-8B	None	46.4	30.7	62.8	34.9	69.4	47.3	54.6	50.0	49.5
	Base	66.1	71.8	82.2	54.7	81.8	82.3	91.9	**65.4**	74.5
	Map	69.5†	**76.1**†	83.4†	**59.6**†	81.7	81.6	92.2	60.5	75.6
	Unit	**71.8**†	74.4	**83.6**†	58.1†	**82.3**	**82.7**	**92.7**†	64.3	**76.2**†
Qwen3-14B	None	51.2	20.0	65.5	36.2	65.4	47.3	81.2	50.0	52.1
	Base	63.2	84.6	87.8	59.5	**83.7**	84.1	93.7	60.5	77.1
	Map	63.4	**85.1**	88.1	63.2†	82.6	83.8	93.6	60.8	77.6
	Unit	**68.4**†	84.9	**88.2**	**66.1**†	83.1	**85.2**	**94.3**	**63.8**†	**79.3**†
Qwen3-32B	None	45.4	32.3	69.8	29.8	62.2	47.3	70.4	49.8	50.9
	Base	58.0	87.5	83.4	69.0	85.7	**83.4**	91.6	70.2	78.6
	Map	60.6†	**92.5**†	83.5	71.1†	85.7	81.2	**92.5**	70.5	79.7†
	Unit	**65.9**†	92.3†	**84.4**	**72.9**†	**85.9**	81.9	92.5	**71.0**	**80.8**†
Mistral-7B-Instruct-v0.1	None	48.6	33.7	62.7	37.4	65.4	49.5	68.3	51.3	52.1
	Base	62.6	39.6	72.4	40.0	65.4	**75.8**	89.2	52.7	62.2
	Map	63.2	36.1	75.1†	43.3†	65.5†	72.9	88.8	54.5	62.4
	Unit	**67.4**†	**40.4**	**75.4**†	**44.9**†	**66.5**†	74.0	**89.4**	**57.5**†	**64.4**†
Mistral-7B-Instruct-v0.3	None	51.3	32.3	63.6	42.1	55.4	56.0	78.4	53.3	54.0
	Base	65.5	**43.0**	80.0	52.4	65.4	77.6	89.7	59.2	66.6
	Map	65.7	40.7	80.7	50.7	**66.4**†	**79.1**	89.7	61.3	66.8
	Unit	**69.7**†	42.5	**81.3**†	**54.2**†	66.0†	78.7	**91.5**†	**62.2**	**68.3**†
Phi-3-medium-4k-instruct	None	25.8	26.8	62.8	30.0	65.4	45.5	50.9	50.0	44.6
	Base	67.0	40.9	82.9	43.1	82.2	**76.5**	**93.8**	55.0	67.7
	Map	**67.6**	43.2	84.4†	47.2†	82.3	65.0	93.3	60.3†	67.9
	Unit	67.6	**43.9**	**85.7**†	**48.5**†	**82.6**	73.3	93.2	**60.8**†	**69.4**†
Phi-3-mini-128k-instruct	None	41.0	25.3	49.8	30.8	65.4	47.3	77.8	50.0	48.4
	Base	62.4	33.3	83.5	49.8	66.6	**78.7**	**91.1**	54.2	65.0
	Map	**64.9**†	38.1†	**83.6**	**54.4**†	**69.3**†	76.5	89.8	**55.6**	**66.5**†
	Unit	64.0	**38.4**†	83.4	53.0†	69.0†	75.8	90.9	55.0	66.2†
Phi-3.5-mini-instruct	None	42.5	25.1	65.1	30.7	65.4	47.3	62.3	50.0	48.6
	Base	61.8	32.5	84.9	54.9	74.3	**81.9**	**91.6**	56.0	67.2
	Map	**66.6**†	34.6	**85.7**†	**59.0**†	**78.0**†	81.6	91.4	**57.7**	**69.3**†
	Unit	64.5†	**38.1**†	85.1	57.8†	77.4†	81.6	91.6	56.9	69.1†
Phi-4	None	41.7	27.2	56.7	34.9	34.6	52.7	50.9	50.0	43.6
	Base	**69.4**	**47.2**	64.3	45.6	68.8	67.9	89.4	**61.0**	64.2
	Map	62.1	35.8	73.3†	41.5	79.9†	62.5	**90.3**	57.4	62.8
	Unit	64.4	45.6	**78.1**†	**50.6**†	**80.0**†	**72.9**	89.7	57.7	**67.4**
MobileLLM-125M	None	39.8	**27.5**	51.8	30.7	36.2	49.1	53.8	49.7	42.3
	Base	65.4	24.2	**62.8**	34.9	**65.4**	52.7	50.9	54.5	51.4
	Map	64.8	26.7	62.8	35.4	65.0	**58.8**	73.2†	56.3	55.4†
	Unit	**69.0**†	24.9	62.8	**38.4**†	65.4	57.8	**75.0**†	**56.9**	**56.3**†
MobileLLM-125M-layer-share	None	37.2	25.8	56.5	30.3	43.2	48.4	49.1	50.5	42.6
	Base	59.1	21.8	**62.8**	**34.7**	**65.4**	52.7	50.9	55.0	50.3
	Map	64.8†	25.6	62.8	34.4	65.4	52.3	54.9†	54.5	51.9†
	Unit	**70.1**†	**26.1**	62.8	34.6	65.3	**55.2**	**66.4**†	**56.9**	**54.7**†
MobileLLM-350M	None	39.6	25.4	60.6	33.9	34.6	53.1	49.1	49.1	43.2
	Base	67.7	23.7	**62.8**	34.7	65.4	50.5	50.9	55.6	51.4
	Map	65.5	**27.4**	62.8	35.4	65.4	**54.9**	72.0†	54.7	54.8†
	Unit	**67.8**	26.3	62.8	**39.0**†	**65.5**	54.2	**73.6**†	**56.4**	**55.7**†
MobileLLM-350M-layer-share	None	34.2	24.7	37.8	30.4	65.4	47.3	50.9	50.0	42.6
	Base	64.7	**24.7**	**62.8**	34.9	65.4	52.7	50.9	49.8	50.8
	Map	70.0†	24.6	62.8	36.6	65.6	56.3	70.0†	**56.9**†	55.3†
	Unit	**71.0**†	23.2	62.8	**38.5**†	**66.4**†	**59.6**†	**71.1**†	54.9†	**55.9**†