DIVERSITY OF TRANSFORMER LAYERS: ONE ASPECT OF PARAMETER SCALING LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers deliver outstanding performance across a wide range of tasks and are now a dominant backbone architecture for large language models (LLMs). Their task-solving performance is improved by increasing parameter size, as shown in the recent studies on parameter scaling laws. Although recent mechanistic-interpretability studies have deepened our understanding of the internal behavior of Transformers by analyzing their residual stream, the relationship between these internal mechanisms and the parameter scaling laws remains unclear. To bridge this gap, we focus on layers and their size, which mainly decide the parameter size of Transformers. For this purpose, we first theoretically investigate the layers within the residual stream through a bias-diversity decomposition. The decomposition separates (i) bias, the error of each layer's output from the ground truth, and (ii) diversity, which indicates how much the outputs of each layer differ from each other. Analyzing Transformers under this theory reveals that performance improves when individual layers make predictions close to the correct answer and remain mutually diverse. We show that diversity becomes especially critical when individual layers' outputs are far from the ground truth. Finally, we introduce an information-theoretic diversity and show our main findings that adding layers enhances performance only when those layers behave differently, i.e., are diverse. We also reveal the performance gains from increasing the number of layers exhibit submodularity: marginal improvements diminish as additional layers increase, mirroring the logarithmic convergence predicted by the parameter scaling laws. Experiments on multiple semantic-understanding tasks with various LLMs empirically confirm the theoretical properties derived in this study. Our code will be available at `https://github.com/[Anonymous]`.

1 INTRODUCTION

Transformer (Vaswani et al., 2017) is one of the dominant architectures across a wide range of tasks in natural language processing (NLP) and other fields. In particular, large language models (LLMs) built on the Transformer backbone have demonstrated remarkable capabilities in language understanding and generation. This rapid progress, however, raises fundamental questions about why Transformers perform so well and how their performance scales with model size.

Empirical scaling studies have shown that increasing model parameter size leads to predictable improvements in performance (Hestness et al., 2017; Kaplan et al., 2020). These scaling laws suggest that larger networks consistently yield better performance as represented by in-context learning ability (Brown et al., 2020) and emergent capabilities observed in models with hundreds of billions of parameters (Wei et al., 2022).

Despite the impressive empirical successes of LLMs, their inner workings largely remain a black box. The mechanistic interpretability has partially elucidated the internal computations of Transformers and uncovered interpretable patterns and circuits within these networks (Geva et al., 2021; Olsson et al., 2022). Studies focusing on the residual stream (Elhage et al., 2021), composed of embedding, Multi-head Attention (MHA), and Multi-layer Perceptron (MLP) layers, have further deepened the interpretation of how information accumulates on residual networks through layers.

These works suggest that each layer contributes incrementally to the model's prediction by adding or refining information in the residual stream. However, the relationship between these internal

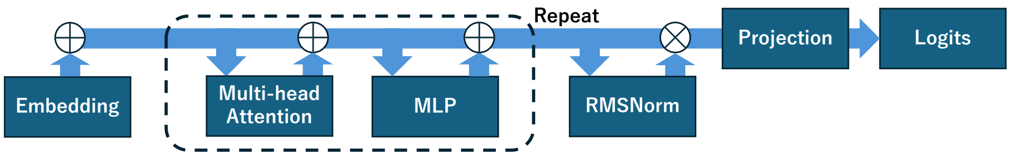

Figure 1: The overview of the residual stream in pre-layer normalization type Transformers.

layer-wise mechanisms and the observed scaling behavior of the overall model remains unclear. This disconnect limits the efficient performance improvement of Transformers.

To address this gap, we propose a theoretical interpretation that links a Transformer's internal layer dynamics with its overall performance, offering a new lens to interpret parameter scaling laws. As a first step, we rely on bias–diversity decomposition (Krogh & Vedelsby, 1994) to quantify the contribution of each layer within the residual stream. In this formulation, each layer's output is evaluated in terms of (i) bias, the error between the layer's output prediction and the ground-truth target, and (ii) diversity, showing how layers' outputs differ from each other. By analyzing Transformers under this approach, we obtain theoretical insights that achieving both low bias and high diversity across layers is ideal for improving performance. In contrast, bias and diversity depend on each other, which makes it difficult to improve both bias and diversity.

Finally, to deal with our main target, parameter scaling laws, from an information-theoretic diversity (Brown, 2009; Zhou & Li, 2010), we examined how diversity contributes to performance gains when additional layers are introduced. Our analysis shows that, to improve performance, the outputs of different layers must remain distinct, i.e., diverse. We also provide a theoretical explanation for the diminishing returns of adding layers: the marginal performance improvement decreases as more layers are stacked. This tendency is consistent with the Scaling Law, which predicts that performance grows logarithmically with the number of parameters.

Finally, we conduct experiments on multiple NLP tasks using various LLMs and show that our theoretical findings are valid in practical situations.

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 $A = (\mathbf{a}^{(1)}, \cdots, \mathbf{a}^{(|A|)})$ be sequences of entire layers, embedding layers, MHA layers, and MLP layers, respectively, the logit prediction on the residual stream is represented as follows:

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

where LN is RMSNorm (Zhang & Sennrich, 2019) and W_{out} is a projection layer. By representing the scaling by RMSNorm as $\mathbf{s}$, we can reformulate Eq. 1 as follows (Chang et al., 2024):

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

Here, by replacing $W_{out}(\mathbf{s} \odot \mathbf{z}^{(i)})$ with $\mathbf{u}^{(i)}$, we can reformulate Eq. 2 as follows:

$$\mathbf{logits} = \sum_{i=1}^{|L|} \mathbf{u}^{(i)}. \tag{3}$$

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 behavior can be formalized as follows:

$$\arg\max_y \mathbf{logits} = \arg\max_y \sum_{i=1}^{|L|} \mathbf{u}^{(i)} = \arg\max_y \frac{1}{|L|} \sum_{i=1}^{|L|} \mathbf{u}^{(i)} \tag{4}$$

Here, we can reformulate Eq. 4 with $\bar{\mathbf{u}} = \frac{1}{|L|} \sum_{i=1}^{|L|} \mathbf{u}^{(i)}$ as follows:

$$(4) \equiv \arg\max_y \bar{\mathbf{u}} \tag{5}$$

By defining $\hat{\mathbf{u}}$ 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):

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) = \frac{1}{|V|} \sum_{j=1}^{|V|} (\hat{u}_j - \bar{u}_j)^2 = \mathbb{E}_{j \in V}[(\hat{u}_j - \bar{u}_j)^2]. \tag{6}$$

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

3.2 IMPORTANCE OF DIVERSITY

First, we focus on the diversity arising from the differences between the predictions of each Transformer layer. From Eqs. 1 and 6, 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:*

$$MSE(\hat{\boldsymbol{u}}, \bar{\boldsymbol{u}}) \tag{7}$$

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

$$= \underbrace{\frac{1}{|L|}\mathbb{E}_{j \in V}[|E| \underbrace{\mathbb{E}_{\mathbf{e} \in E}[(\hat{u}_j - e_j)^2]}_{Embedding\ Bias} + |M| \underbrace{\mathbb{E}_{\mathbf{m} \in M}[(\hat{u}_j - m_j)^2]}_{MLP\ Bias} + |A| \underbrace{\mathbb{E}_{\mathbf{a} \in A}[(\hat{u}_j - a_j)^2]}_{Attention\ Bias}}_{\text{Bias}} \tag{9}$$

$$- \underbrace{\frac{1}{|L|}\mathbb{E}_{j \in V}[|E| \underbrace{\mathbb{E}_{\mathbf{e} \in E}[(\bar{u}_j - e_j)^2]}_{Embedding\ Diversity} + |M| \underbrace{\mathbb{E}_{\mathbf{m} \in M}[(\bar{u}_j - m_j)^2]}_{MLP\ Diversity} + |A| \underbrace{\mathbb{E}_{\mathbf{a} \in A}[(\bar{u}_j - a_j)^2]}_{Attention\ Diversity}}_{\text{Diversity}}. \tag{10}$$

(See Appendix A.1 for the proof.)

As shown in Eq. 8, 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 9 and 10 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, while in contrast, its relative importance may increase in lightweight models. We empirically investigate these hypotheses through experiments in §5.3.

3.3 Limitation of Diversity

As discussed in §3.2, the diversity term is important to improve the prediction performance. However, there are limitations to its effectiveness. The next theorem indicates the limitation.

Theorem 2 (**Limitation of Diversity**). *The decomposition of the prediction error in Transformer layers (Eqs. 8, 9, and 10) holds the following relation:* $Diversity \to 0$ $(Bias \to 0)$*;* $Embedding\ Diversity \to 0$ $(Embedding\ Bias \to 0)$*;* $MLP\ Diversity \to 0$ $(MLP\ Bias \to 0)$*;* $Attention\ Diversity \to 0$ $(Attention\ Bias \to 0)$*. (See Appendix A.2 for the proof.)*

From Theorem 2, regarding the prediction by entire modules and each module, when bias is close to zero, we cannot expect performance improvement from diversity. Therefore, when the predictions of each layer are sufficiently accurate, performance improvement by increasing diversity is limited. However, this relationship does not guarantee that diversity will increase when bias is far from zero. Therefore, when the predictions of each layer are inaccurate, the importance of diversity increases. Even under this situation, the following limitation exists.

Theorem 3 (**Bias and Diversity Trade-off**). *Bias and Diversity in the decomposition of the prediction error in Transformer layers (Eq. 8) depend on each other as follows (Brown et al., 2005):*

$$Bias = \overline{bias}^2 + \Omega,\ Diversity = \Omega - \left[\frac{1}{|L|}\overline{var} + \left(1 - \frac{1}{|L|}\right)\overline{cov}\right], \tag{11}$$

where $\overline{bias} = \mathbb{E}_{\mathbf{u}\in L}[\mathbb{E}_{j\in V}[u_j] - \mathbb{E}_{j\in V}[\hat{u}_j]]$, $\overline{var} = \mathbb{E}_{\mathbf{u}\in L}[\mathbb{E}_{j\in V}(u_j - \mathbb{E}_{k\in V}[u_k])^2]$, $\overline{cov} = \frac{1}{|L|(|L|-1)}\sum_{i=1}^{|L|}\sum_{i\neq j}^{|L|}\mathbb{E}_{k\in V}[(u_k^i - \mathbb{E}_{l\in V}[u_l^i])(u_k^j - \mathbb{E}_{l\in V}[u_l^j])]$*, and* $\Omega = \overline{var} + \mathbb{E}_{\mathbf{u}\in L}[(\mathbb{E}_{j\in V}[u_j] - \mathbb{E}_{j\in V}[\overline{u}_j])^2]$ *(See Appendix A.3 for the proof.)*

By focusing on Eq. 11, 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 3 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.

3.4 Limitation of Stacking Layers

In §3.3, we discussed a trade-off relationship between the bias and diversity terms, and improving the diversity term can potentially improve the prediction performance of Transformers when the bias term is large. A simple method to improve performance is to add layers following the parameter scaling laws, and it can also improve the diversity based on our analysis so far. We theoretically investigate the performance improvement and limitations of this method through the following theorems.

Theorem 4 (**Generalized Diveristy**). *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*[1] *predicting the best label. Its prediction error* $p(Y \neq g(U_{1:|L|}))$ *satisfies the following inequality (Brown, 2009; Zhou & Li, 2010):*

$$\frac{H(Y) - I(U_{1:|L|};Y) - 1}{\log|Y|} \leq p(Y \neq g(U_{1:|L|})) \leq \frac{H(Y) - I(U_{1:|L|};Y)}{2}, \tag{12}$$

[1]This function must follow the Bayes decision rule according to 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).

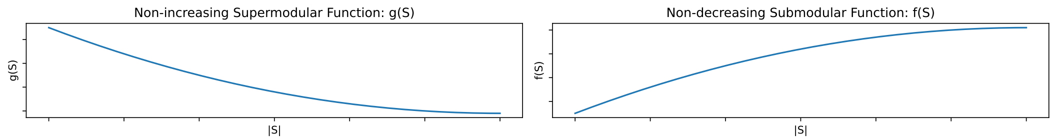

Figure 2: The examples of non-increasing supermodular and non-decreasing submodular functions.

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):

$$I(U_{1:|L|};Y) = \underbrace{\sum_{i=1}^{|L|} I(\mathbf{u}^{(i)};Y)}_{Relevancy} + \underbrace{\underbrace{\mathcal{I}(U_{1:|L|}|Y)}_{Conditional\ Redundancy} - \underbrace{\mathcal{I}(U_{1:|L|})}_{Redundancy}}_{Information\ Theoretic\ Diveristy}, \tag{13}$$

where $\mathcal{I}(U_{1:|L|})$ is total correlation and $\mathcal{I}(U_{1:|L|};Y)$ is conditional total correlation. (See Appendix A.4 for the proof.)

We can interpret Theorem 4 from the bias and diversity decomposition in Eq. 10 from a viewpoint of information theory, indicating that the bias term corresponds to the relevancy term and the diversity term corresponds to the information-theoretic diversity. Through this decomposition, we discuss how adding layers improves the prediction performance of Transformers.

Theorem 5 (**Monotonicity of each Term**). *In Eq. 13, when $U_{1:|L|}$ increases (that means new elements are added to $U_{1:|L|}$), Relevancy, Conditional Redundancy, and Redundancy monotonically increase, whereas Information Theoretic Diveristy and $I(U_{1:|L|};Y)$ do not monotonically increase or decrease (See Appendix A.5 for the proof.)*

Theorem 5 suggests that simply adding more layers to a Transformer does not guarantee an improvement in performance, especially due to the difficulty of improving information-theoretic diversity. This insight is further reinforced by the following theorem.

Theorem 6 (**Monotonicity of Upper and Lower Bounds**). *In Eq. 12, the lower bound $\frac{H(Y)-I(U_{1:|L|};Y)-1}{\log|Y|}$ and the upper bound $\frac{H(Y)-I(U_{1:|L|};Y)}{2}$ do not monotonically increase or decrease when $U_{1:|L|}$ increases. (See Appendix A.6 for the proof.)*

These results seem to indicate that increasing the number of layers in Transformers makes it difficult to improve performance, but at the same time, they suggest that performance can be improved by enhancing diversity corresponding to *Information Theoretic Diveristy*. Therefore, similar to Theorem 1, diversity is also important from an information-theoretical perspective. However, regarding the information-theoretic diversity, there is a limitation described by the following theorem.

Theorem 7 (**Submodularity of each Term**). *In Eq. 13, 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 A.7 for the proof.)*

This theorem assumes a kind of lower bound that *Conditional Redundancy* becomes zero due to the variables in $U_{1:|L|}$ being independent given Y. Even in this situation, the performance of Transformers increases by adding layers because $I(U_{1:|L|};Y)$ is non-decreasing. Here, submodularity refers to the property whereby an increase in input leads to decreasing additional benefit in output. Thus, the effect of adding layers on performance decreases as the number of layers increases in this situation. By assuming that the number of layers and the number of parameters are proportional, this result is consistent with the parameter scaling laws, which show that performance converges logarithmically with an increase in the number of parameters. The next theorem reinforces this characteristic.

Theorem 8 (**Supermodularity of Upper and Lower Bounds**). *In Eq. 12, 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 A.8 for the proof.)*

Supermodularity is a property that the more inputs there are, the greater the risk that additional inputs will result in performance not improving. Therefore, this theorem is also consistent with the parameter scaling laws and reinforces Theorem 7.

Figure 2 shows the example behaviors of non-increasing supermodular and non-decreasing submodular functions. The gradual convergence of values aligns with the logarithmic convergence of parameter scaling laws. However, these are based on a kind of lower bound that *Conditional Redundancy* becomes zero. In the actual situation, we can expect performance improvement by increased *Conditional Redundancy*, and it makes the performance improvement non-monotonic, by following Theorems 5 and 6. We check the actual behavior in §5.5.

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. 8 using the following equation:

$$MSE(\hat{\mathbf{u}}, \bar{\mathbf{u}}) \approx MSE(\tilde{\mathbf{u}}, softmax(\bar{\mathbf{u}})); bias \approx \frac{1}{|V|} \sum_{i=1}^{|V|} \frac{1}{|L|} \sum_{j=1}^{|L|} \left(\widetilde{u}_i - softmax(\mathbf{u}^{(j)})_i \right)^2, \quad (14)$$

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

4.2 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, this assumption overlooks several important factors. One such factor is the number of dimensions. In many models, the actual number of parameters is represented by the product of the number of dimensions and the number of layers. Therefore, 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.

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 of such a model 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

5.1 General Settings

We used the following Transformer models from HuggingFace Transformers (Wolf et al., 2020): facebook/MobileLLM-125M; facebook/MobileLLM-125M-layer-share; facebook/MobileLLM-350M; facebook/MobileLLM-350M-layer-share (Liu et al., 2024); meta-llama/Llama-2-7b-hf;

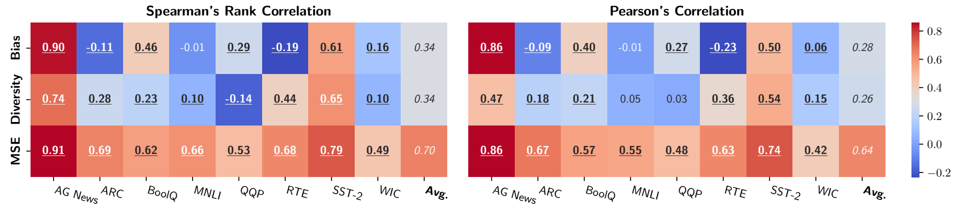

Figure 3: The correlation between accuracy and MSE, bias, and diversity of models across tasks. The underlined scores show the statistical significance ($p < 0.05$).[3]Note that the scores on Avg. are not the target of the significance test.

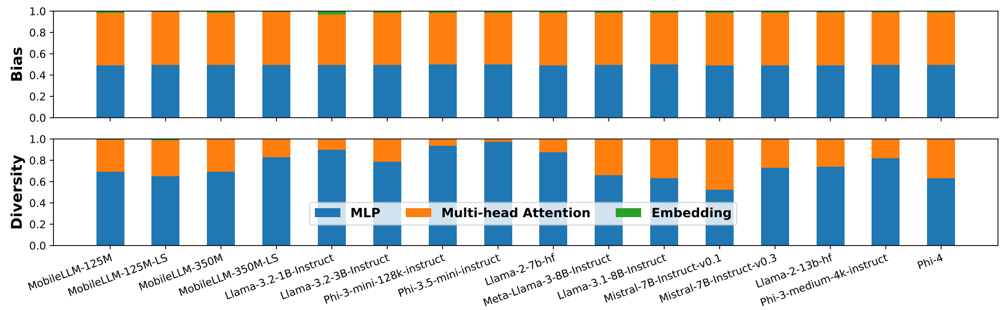

Figure 4: The average proportion of bias and diversity accounted for by each module in each model in all datasets.

meta-llama/Llama-2-13b-hf (Touvron et al., 2023); meta-llama/Meta-Llama-3-8B-Instruct; meta-llama/Llama-3.1-8B-Instruct; meta-llama/Llama-3.2-1B-Instruct; meta-llama/Llama-3.2-3B-Instruct Grattafiori et al. (2024); microsoft/Phi-3-medium-4k-instruct; microsoft/Phi-3-mini-128k-instruct; microsoft/Phi-3.5-mini-instruct Abdin et al. (2024a); microsoft/phi-4 (Abdin et al., 2024b); mistralai/Mistral-7B-Instruct-v0.1; mistralai/Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).

By following the setting of the previous work (Chang et al., 2024), we used the first 2000 instances in 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., 2019) datasets. We use the accuracy as a metric for each task. We used a predefined template by for prompting on each task and restricted the output to predefined options (See Appendix B for the details).

5.2 Correlation of MSE, Bias, and Diversity to Performance

In order to clarify the relationship between MSE, bias, diversity, and performance in actual tasks, we calculated the Pearson and Spearman's correlation between the accuracy and MSE, bias, and diversity at each junction point on the residual stream. Figure 3 shows the correlation.[4] As can be seen from these results, MSE shows a high correlation with accuracy. On the other hand, bias and diversity show moderate correlations, but the trends differ depending on the task, indicating that the model changes the roles of each layer according to the task. Furthermore, the fact that the Spearman correlation is lower than the rank correlation indicates that detailed accuracy differences cannot be read from MSE.

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

[4]Different from accuracy and diversity, lower bias and MSE are better for performance improvement. To reflect it, we used the negative value of bias and MSE when calculating the correlation. Also, to assign the value range, we standardize each score for each model. By utilizing z-transformation (Corey et al., 1998), we report averaged correlation (Avg.).

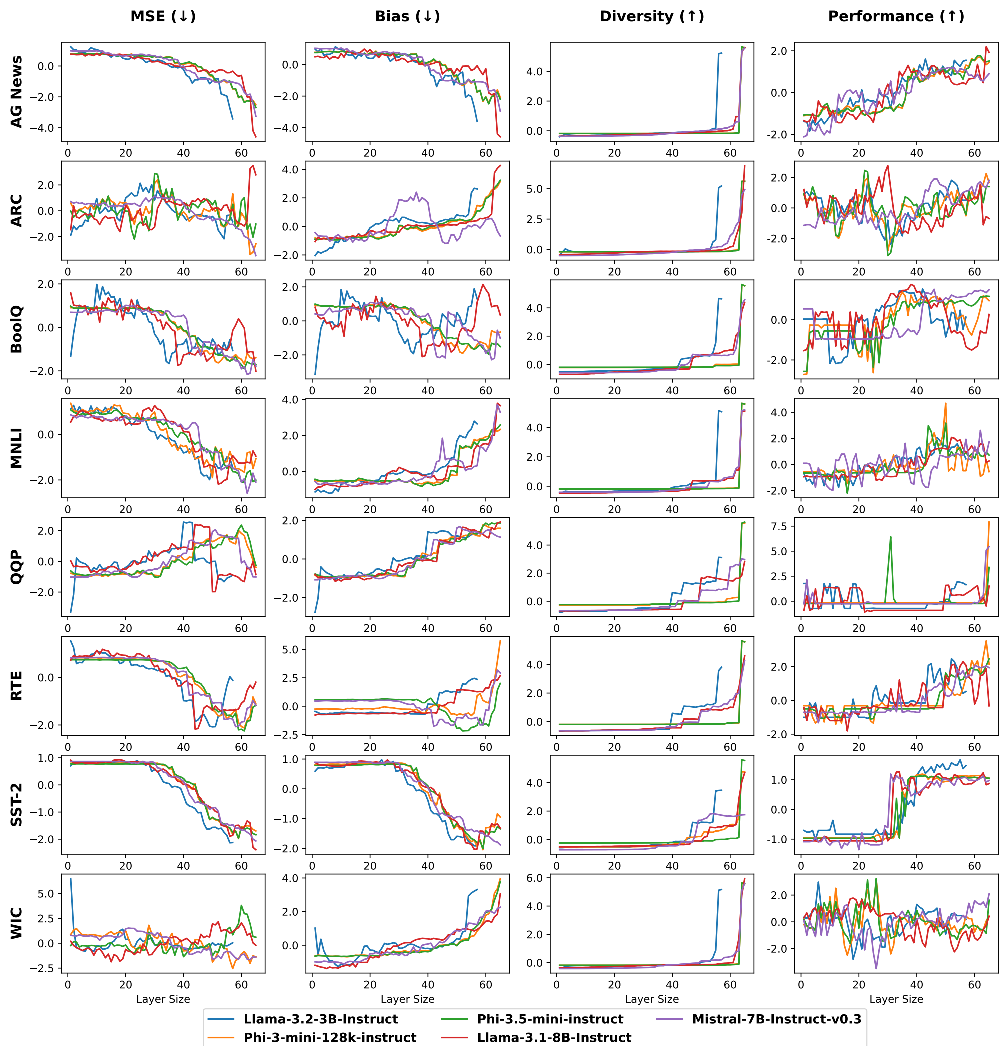

Figure 5: The relationship between MSE, bias, diversity, and performance on each task. Note that each metric is standardized to capture the change by the number of layers.

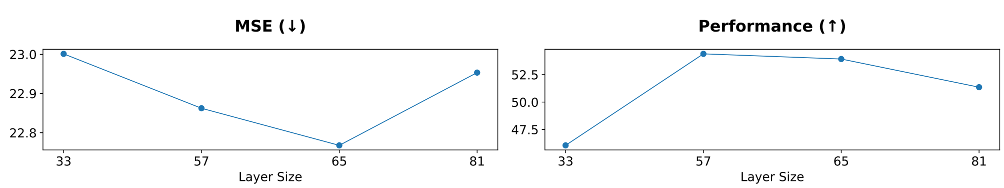

Figure 6: The relationship between performance and the number of layers.

5.3 Influence of each Module

Figure 4 shows the influence of each module on bias and diversity based on the weights corresponding to their layer size, represented by Eqs. 9 and 10. 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

Table 1: The results of MobileLLM on different tasks. Avg. indicates the average scores of all tasks. The suffix -LS indicates that the model recurrently shares its layers.

MobileLLM	Accuracy									MSE								
	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									Diversity								
	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

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.

5.4 Trade-off Relationship between Bias and Diversity

We investigate whether there is actually a trade-off between bias and diversity, as shown in Theorems 2 and 3. Figure 5 shows the results.[5] As can be seen from these results, when there is no performance improvement, there is a trade-off between bias and diversity, but when performance improves, the trade-off relationship does not hold. This is consistent with the fact that bias and diversity share only one term, as shown in Theorem 3. The limitation of diversity shown in Theorem 2 is caused by bias being close to zero, but in reality, bias rarely approaches zero, so 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.

5.5 Decrease in Efficiency with Increasing Number of Layers

We verify whether the decrease in performance improvement accompanying the increase in layers shown in Theorems 5, 6, 7, and 8 actually occurs. To do so, we averaged the accuracy and MSE for all tasks in models with no less than 1B of parameters. Figure 6 shows the result. From these results, we can see that the effectiveness of increasing layers actually decreases. Also, the performance degradation caused by increasing layers indicates that *Conditional Redundancy* becomes larger than zero, and the layers depend on each other to predict the answers.

5.6 Consistency of Our Theory with Models that Reuse Layers

We verify the performance of a model that reuses layers. Table 1 shows the results obtained using MobileLLM. Basically, we find that the performance improvement achieved by reusing layers is limited. Furthermore, since the diversity of models that reuse layers is generally low, we consider that the low diversity is the reason for the limited performance improvement. This suggests that in order to improve model performance by reusing layers, it is necessary to improve the diversity of each module.

6 Conclusion

In this paper, we demonstrated that the diversity of predictions from each module on the residual stream of Transformers is important for improving performance from the perspective of bias and diversity decomposition. Additionally, we demonstrated limits to the performance improvements achievable through this diversity. Furthermore, we showed that performance improvements achieved by adding more layers are related to this diversity and that the effectiveness of these improvements diminishes as the number of layers increases, which is consistent with the parameter scaling laws. Experimental results across multiple tasks with various LLMs confirmed that empirical observations support these theoretical claims.

[5] Note that due to low visibility, this figure only covers models that achieved the top five accuracy on average over all tasks. See Appendix C for the overall result.

Ethics Statement

Because our work uses predefined output and templates for the investigation, there is no possibility of generating harmful content from LLMs. Furthermore, our work can contribute to revealing the internal behavior of LLMs to understand given instructions, which is often useful to prevent LLMs from generating harmful content.

Reproducibility Statement

Experimental settings and their details are described in §5 and Appendix B, respectively. Both sections cover the essential information for reproducing our reported results, which are about the used templates, models, decoding methods, and GPU environment. In addition, we will release our code for 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.

Rishabh Iyer, Ninad Khargoankar, Jeff Bilmes, and Himanshu Asanani. Submodular combinatorial information measures with applications in machine learning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 722–754. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/iyer21a.html.

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.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn

Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

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. Probable error of a correlation coefficient. *Biometrika*, pp. 302–310, 1908.

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

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 R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019. In the Proceedings of ICLR.

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.

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 Proofs

A.1 The proof for Theorem 1

According to Krogh & Vedelsby (1994), when the coefficients of each term in Eq. 6 satisfy the definition of probability, we can decompose Eq. 6 into Eq. 8. Here, the coefficients follow a uniform distribution, and the decomposition of Eq. 8 holds. We can induce Eqs. 9 and 10 based on the fact that the weighted mean equals the original mean.

A.2 The proof for Theorem 2

When the bias term in Eq. 8 equals zero, $\hat{u} = u$ holds. In this condition, $\bar{u} = \sum u$ becomes $\hat{u}$. Thus, the diversity terms become zero. Similarly, other decomposed diversity terms become zero when corresponding bias terms become zero.

A.3 The proof for Theorem 3

The decomposition of Krogh & Vedelsby (1994) is a sufficient condition for the decomposition of Brown et al. (2005). Since we proved that Theorem 2 holds, Theorem 3 also holds.

A.4 The proof for Theorem 4

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.

A.5 The proof for Theorem 5

Since *Relevancy* is a sum of joint mutual information, *Relevancy* increases when $|U_{1:|L|}|$ increases. This is because joint mutual information always becomes a positive value.

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:

$$
\begin{aligned}
&\mathcal{I}(U_{1:|L|+1}|Y) - \mathcal{I}(U_{1:|L|}|Y) && (15)\\
=&\left(\sum_{i=1}^{|L|+1} H(\mathbf{u}^{(i)}|Y)\right) - H(U_{1:|L|+1}|Y) - \left(\sum_{i=1}^{|L|} H(\mathbf{u}^{(i)}|Y)\right) + H(U_{1:|L|}|Y) && (16)\\
=&H(\mathbf{u}^{(|L|+1)}|Y) - H(U_{1:|L|+1}|Y) + H(U_{1:|L|}|Y) && (17)\\
=&H(\mathbf{u}^{(|L|+1)}|Y) - \Big(H(\mathbf{u}^{(|L|+1)}|Y) + H(U_{1:|L|}|Y, \mathbf{u}^{(|L|+1)})\Big) + H(U_{1:|L|}|Y) && (18)\\
=&H(U_{1:|L|}|Y) - H(U_{1:|L|}|Y, \mathbf{u}^{(|L|+1)}) && (19)\\
=&I(U_{1:|L|+1}|Y). && (20)
\end{aligned}
$$

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:

$$\begin{aligned}
&\mathcal{I}(U_{1:|L|+1}) - \mathcal{I}(U_{1:|L|}) && (21)\\
=&\left(\sum_{i=1}^{|L|+1} H(\mathbf{u}^{(i)})\right) - H(U_{1:|L|+1}) - \left(\sum_{i=1}^{|L|} H(\mathbf{u}^{(i)})\right) + H(U_{1:|L|}) && (22)\\
=&H(\mathbf{u}^{(|L|+1)}) - H(U_{1:|L|+1}) + H(U_{1:|L|}) && (23)\\
=&H(\mathbf{u}^{(|L|+1)}) - \Big(H(\mathbf{u}^{(|L|+1)}) + H(\mathbf{u}^{(|L|+1)}|U_{1:|L|})\Big) + H(U_{1:|L|}) && (24)\\
=&H(U_{1:|L|}) - H(\mathbf{u}^{(|L|+1)}|U_{1:|L|}) && (25)\\
=&I(U_{1:|L|+1}). && (26)
\end{aligned}$$

Since $I(U_{1:|L|+1})$ is non-negative, *Redundancy* monotonically increases. Note that 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. 20 and 26 as follows:

$$\begin{aligned}
&\mathcal{I}(U_{1:|L|+1}|Y) - \mathcal{I}(U_{1:|L|}|Y) - (\mathcal{I}(U_{1:|L|+1}) - \mathcal{I}(U_{1:|L|})) && (27)\\
=&I(U_{1:|L|+1}|Y) - I(U_{1:|L|+1}). && (28)
\end{aligned}$$

When elements in $U_{1:|L|+1}$ are independent but not independent given Y, Eqs. 28 is non-negative. On the other hand, when elements in $U_{1:|L|+1}$ are not independent but independent given Y, Eq. 28 is non-positive. Thus, *Information Theoretic Diversity* does not have monotonicity.

Finally, Iyer et al. (2021) indicates the non-monotonicity of mutual information. Therefore, $I(U_{1:|L|};Y)$ does not satisfy the monotonicity. We can check its increase when $U_{1:|L|}$ obtains a new element, $\mathbf{u}^{(|L|+1)}$ as follows:

$$I(\mathbf{u}^{(|L|+1)};Y) + I(U_{1:|L|+1}|Y) - I(U_{1:|L|+1}). \tag{29}$$

When elements in $U_{1:|L|+1}$ are independent but not independent given Y, Eq. 29 is non-negative. On the other hand, when elements in $U_{1:|L|+1}$ are not independent but independent given Y, Eq. 29 is non-positive.

A.6 THE PROOF FOR THEOREM 6

Only $I(U_{1:|L|};Y)$ is a term including $U_{1:|L|}$ in both bounds, $I(U_{1:|L|};Y)$ dominates the bounds. Because $I(U_{1:|L|};Y)$ is not a monotonic function, the upper and lower bounds do not satisfy the monotonicity.

A.7 THE PROOF FOR THEOREM 7

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

$$\begin{aligned}
\Delta(\mathbf{u}|U_{1:|L|}) =&I(U_{1:|L|}, \mathbf{u}^{(|L|+1)};Y) - I(U_{1:|L|};Y) && (30)\\
=&I(Y,\mathbf{u}|U_{1:|L|}) + I(U_{1:|L|};Y) - I(U_{1:|L|};Y) && (31)\\
=&I(Y,\mathbf{u}|U_{1:|L|}) && (32)
\end{aligned}$$

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,

$$I(Y,\mathbf{u}|U_{1:|L|}) = H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|Y,U_{1:|L|}), \tag{33}$$

Table 2: 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

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

$$\begin{aligned}
&\Delta(\mathbf{u}|U_{1:|L|}) - \Delta(\mathbf{u}|U_{1:|L|+1}) && (34)\\
=&H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|Y, U_{1:|L|}) - \big(H(\mathbf{u}|U_{1:|L|+1}) - H(\mathbf{u}|Y, U_{1:|L|+1})\big) && (35)\\
=&H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|U_{1:|L|+1}) - \big(H(\mathbf{u}|Y, U_{1:|L|}) - H(\mathbf{u}|Y, U_{1:|L|+1})\big) && (36)\\
=&H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|U_{1:|L|+1}) - (H(\mathbf{u}|Y) - H(\mathbf{u}|Y)) && (37)\\
=&H(\mathbf{u}|U_{1:|L|}) - H(\mathbf{u}|U_{1:|L|+1}) && (38)\\
\geq&0 && (39)
\end{aligned}$$

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 Fujishige (1978); Studeny (2006). Since $\mathcal{I}(U_{1:|L|}|Y)$ is zero under the given condition, *Information Theoretic Diveristy* becomes

$$-\mathcal{I}(U_{1:|L|}). \tag{40}$$

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

A.8 The proof for Theorem 8

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.

B Experimental Details

Table 2 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 3 shows the templates used for each task. All templates are extracted from promptsource (Bach et al., 2022)[6]. The output vocabulary is restricted to the tokens used in options corresponding to the task. Table 4 shows the details of the used models in our experiments. In the implementation, we modified the released code from Chang et al. (2024)[7]. We used an NVIDIA RTX A6000, which has 48GB of VRAM, to run the models.

C Detailed Results

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

[6] https://github.com/bigscience-workshop/promptsource
[7] https://github.com/terarachang/LLMDecomp

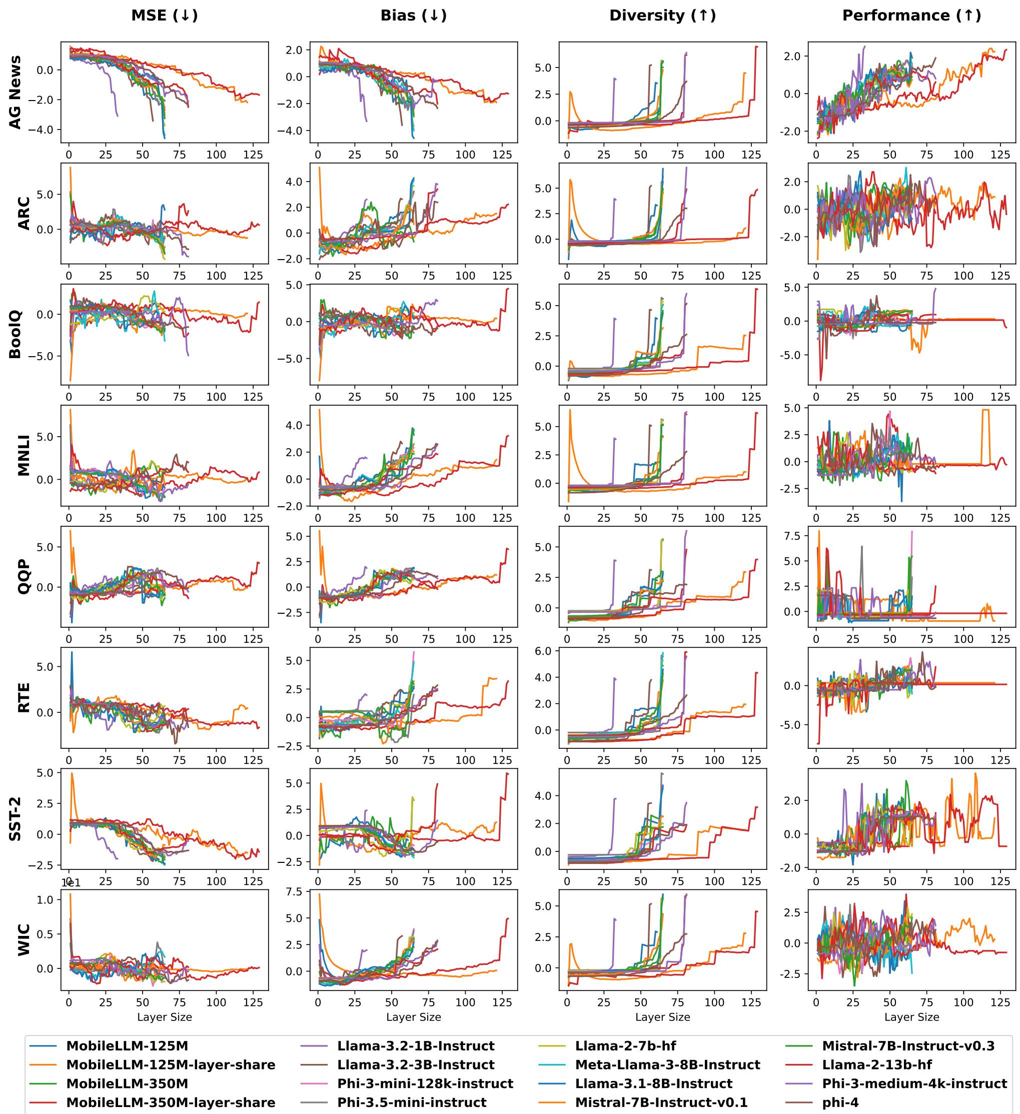

Figure 7: The relationship between MSE, bias, diversity, and performance of all models on each task. The settings are the same as Figure 5.

Table 3: 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

Table 4: The statistics of the used models.

Model Name	Parameter Size	Layer Size	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
meta-llama/Llama-3.2-1B-Instruct	1B	33	LLaMA License
meta-llama/Llama-3.2-3B-Instruct	3B	57	LLaMA 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 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 License
meta-llama/Llama-3.1-8B-Instruct	8B	65	LLaMA License
meta-llama/Llama-2-13b-hf	13B	81	LLaMA License
microsoft/Phi-3-medium-4k-instruct	14B	81	MIT
microsoft/phi-4	14.7B	81	MIT

D LLM USAGE

We used GPT-4o (OpenAI et al., 2024) only for proofreading our paper.