Scaling Law Analysis in Federated Learning: How to Select the Optimal Model Size?

Xuanyu Chen[1], **Nan Yang**[1*], **Shuai Wang**[2], **Dong Yuan**[1*],

[1]The University of Sydney, Sydney, Australia
[2]Northwestern Polytechnical University, Xi'an, China
xuanyu.chen@sydney.edu.au, n.yang@sydney.edu.au, s.wang@mail.nwpu.edu.cn, dong.yuan@sydney.edu.au

Abstract

The recent success of large language models (LLMs) has sparked a growing interest in training large-scale models. As the model size continues to scale, concerns are growing about the depletion of high-quality, well-curated training data. This has led practitioners to explore training approaches like Federated Learning (FL), which can leverage the abundant data on edge devices while maintaining privacy. However, the decentralization of training datasets in FL introduces challenges to scaling large models, a topic that remains under-explored. This paper fills this gap and provides qualitative insights on generalizing the previous model scaling experience to federated learning scenarios. Specifically, we derive a PAC-Bayes (Probably Approximately Correct Bayesian) upper bound for the generalization error of models trained with stochastic algorithms in federated settings and quantify the impact of distributed training data on the optimal model size by finding the analytic solution of model size that minimizes this bound. Our theoretical results demonstrate that the optimal model size has a negative power law relationship with the number of clients if the total training compute is unchanged. Besides, we also find that switching to FL with the same training compute will inevitably reduce the upper bound of generalization performance that the model can achieve through training, and that estimating the optimal model size in federated scenarios should depend on the average training compute across clients. Furthermore, we also empirically validate the correctness of our results with extensive training runs on different models, network settings, and datasets.

Extended Version — https://arxiv.org/abs/2511.12188

Introduction

In recent years, large-scale models with billions of parameters have been introduced and shown impressive performance on many tasks, such as Large Language Models (LLMs) (Brown et al. 2020; Rae et al. 2021; Thoppilan et al. 2022). Given the high cost of repeatedly training these massive models to achieve optimal results (Zhai et al. 2022; Tay et al. 2021), researchers have developed scaling laws to estimate the compute-optimal model size, minimizing the need for repeated training, revealing that larger models can achieve better performance if trained with more data, as prescribed by these laws. For example, the Chinchilla model (Hoffmann et al. 2022), with 70 billion parameters, outperformed the Gopher model (Rae et al. 2021), which has 280 billion parameters, by following these laws to be trained on four times more data.

However, significantly expanding high-quality training data on centralized servers has long been a challenge (Muennighoff et al. 2023). In contrast, vast amounts of data are generated and stored in a distributed manner across society every day. A popular technique for utilizing this distributed data is Federated Learning (FL) (McMahan et al. 2017; AbdulRahman et al. 2021), where multiple clients collaboratively train a model while maintaining data privacy by preserving the training data locally. Notably, although it is less challenging to increase the size of training data in a federated learning scenario by scaling the network size (increasing the number of clients), the decentralization of training data will undoubtedly cause an impact on the scaling behavior of large-scale models summarized in previous studies. This motivates a question for practitioners seeking optimal training results: ***How will the estimation of the optimal model size be affected when training large-scale models with distributed data using Federated Learning?***

In this paper, we provide qualitative insights into generalizing existing model scaling laws to federated learning scenarios, supported by both theoretical analysis and empirical evidence. Theoretically, we model the Federated Learning process as a Stochastic Gradient Descent (SGD) optimization problem (Bottou 1998; Sutskever et al. 2013) over distributed data and establish a PAC-Bayes upper bound (McAllester 1998, 1999) for the generalization error of this learning algorithm. Using this generalization bound, we derive an analytic solution of the optimal model size minimizing this bound, quantifying the impact of distributed data on this optimal size. Empirically, we conduct several experiments to validate our theoretical findings. We pre-trained numerous models with varying parameter sizes using a popular architecture of auto-regressive transformer (Vaswani et al. 2017), Vision Transformer (Dosovitskiy et al. 2021; He et al. 2022) under centralized and federated settings. We then conducted linear probing on these models using two standard datasets, CIFAR-100 (Krizhevsky and Hinton 2009) and ImageNet (Deng et al. 2009), and analyzed the final loss to

*Nan Yang and Dong Yuan are Corresponding Authors.

demonstrate variations in optimal model size, which closely align with our theoretical predictions.

In summary, the main contributions of this paper are shown below:

1. We prove a PAC-Bayes upper bound for the generalization error of models trained by stochastic gradient descent optimization under federated settings and derive the analytic solution of the optimal model size under the convex settings of this bound.
2. By analyzing the analytic solution of the optimal model size, we obtain several qualitative insights on generalizing the scaling experience to federated scenarios: (i) the optimal model size has a negative power law relationship with the number of clients if the total training compute is unchanged; (ii) data decentralization in federated scenarios with a large number of clients will cause a drop in the upper bound of generalization performance that models can achieve through training; (iii) the optimal model size in federated scenarios can also be estimated from the average training compute allocated to each client.
3. Furthermore, we conducted a sufficient number of training runs on models with different parameter sizes through centralized learning and federated learning, respectively. The empirical evidence validates the correctness of our theoretical results.

Related Work

Scaling Law of LLMs. Given the infeasibility of repeatedly training large language models (LLMs) with billions of parameters (Zhai et al. 2022; Tay et al. 2021), researchers have developed various scaling laws to predict the relationship between the optimal model size and available training resources. Kaplan et al. were the first to observe a power-law relationship between model performance and model size (Kaplan et al. 2020), laying the foundation for subsequent works. Hoffman et al. revisited the problem under computational constraints and proposed the Chinchilla scaling law, which recommends equally scaling both model size and dataset size (Hoffmann et al. 2022). Recent studies have noted that the Chinchilla scaling law could deplete available training data, leading to the development of new scaling laws for data-constrained scenarios (Gadre et al. 2024; Muennighoff et al. 2023). However, these scaling laws are derived from empirical results in centralized training and may not directly apply to scenarios where training data is distributed. Our work addresses this gap by providing a theoretical analysis of the generalization bound and empirically validating the impact of data decentralization on scaling laws.

Federated Training of Large-scale Models. Federated Learning (FL) has garnered significant attention for enabling collaborative model training while preserving data privacy by keeping local data on clients (McMahan et al. 2017; AbdulRahman et al. 2021; Tang et al. 2023). Specifically, clients receive the global model from the central server, compute updates using their local data, and send these updates back to the server. The server aggregates the updates to refine the global model. This process is repeated until the model achieves the desired performance. With the rise of Large-scale models like LLMs, there has been a growing interest in applying FL to train these large-scale models (Chen et al. 2023). Since clients generally have fewer computational resources than the server, most of these works follow the intuition to reduce the model size by tailoring the architecture of language models or freezing part of the model parameters during training (Xu et al. 2024; Wu, Liang, and Wang 2020). However, few studies have explored the modified scaling behavior of language models in federated scenarios (Hilmkil et al. 2021; Ro et al. 2022; Shen et al. 2025), and those that have primarily offered observational insights based on empirical evidence. To address this theoretical gap, we model federated training as an SGD optimization problem over distributed data and quantify the impact on scaling by deriving an analytic solution for the optimal model size.

Generalization Bound for Stochastic Algorithms. Stochastic gradient descent (SGD) (Bottou 1998; Sutskever et al. 2013) is a widely used optimization method in machine learning (LeCun et al. 1998; Hinton and Salakhutdinov 2006; Goodfellow et al. 2014; McMahan et al. 2017; Tang et al. 2023). Previous research has shown that the generalization performance of stochastic algorithms can be quantified using a PAC-Bayes upper bound (He, Liu, and Tao 2019; Mou et al. 2018; London 2017; Pensia, Jog, and Loh 2018), which is applied to explore various aspects, including algorithm convergence (Mou et al. 2018; Pensia, Jog, and Loh 2018), training stability (Zhu et al. 2024), or strategy of tuning hyper-parameters (He, Liu, and Tao 2019). The generalization bound also provides treatment for federated learning, helping several studies to propose new training frameworks to address the non-IID problem (Zhao et al. 2024) or model personalization (Boroujeni, Krause, and Ferrari Trecate 2024; Achituve et al. 2021; Vedadi et al. 2024), and other studies to figure out the impact of common parameters in federated scenarios on training results (Sefidgaran et al. 2024). In contrast, instead of proving similar generalization bounds for one of the two training regimes, our work focuses on comparing the generalization bounds of the stochastic algorithms in the federated settings with those in the centralized settings. The comparison results show us the impact of changing the training scenario on the optimal model size.

Preliminaries

Bound for Generalization Error

Formally, considering the hypothesis class of a model is $\Theta \subset \mathbb{R}^d$, machine learning algorithms aim to find the vector of model parameters $\theta \in \Theta$ that minimizes the expected risk $\mathcal{R}(\theta) = \mathbb{E}_{\xi\sim\mathcal{D}} F(\theta;\xi)$ where d is the dimension of the parameter θ, F is the loss function, and $\mathcal{D}$ is the latent distribution of testing data. Suppose the output parameter θ follows a distribution Q, the expected risk in terms of Q can be formulated as:

$$\mathcal{R}(Q) = \mathbb{E}_{\theta\sim Q}\mathbb{E}_{\xi\sim\mathcal{D}} F(\theta;\xi). \tag{1}$$

In practice, since $\mathcal{D}$ is not known in advance, the expected risk $\mathcal{R}$ is estimated by the empirical risk $\hat{\mathcal{R}}$ in terms of the latent distribution $\hat{\mathcal{D}}$ of the training data and is defined as:

$$\hat{\mathcal{R}}(Q) = \mathbb{E}_{\theta \sim Q} \mathbb{E}_{\zeta \sim \hat{\mathcal{D}}} F(\theta; \zeta). \tag{2}$$

The difference between $\mathcal{R}$ and $\hat{\mathcal{R}}$ is known as a generalization error, and the upper bound of the generalization error is usually used as a critical index to demonstrate the generalization ability of the training algorithm.

SGD Optimization

Stochastic Gradient Descent (SGD) is typically used to optimize the empirical risk $\hat{R}$. Consider a training dataset with size m, the mini-batch $\mathcal{S}$ of the training samples is equivalent to a subset of S random indices that are independently and identically (i.i.d.) drawn from the index set $\{1, \dots, m\}$. The SGD iteration can be formally defined as:

$$\begin{aligned} \theta(t+1) &= \theta(t) - \eta \nabla_{\theta(t)} \hat{\mathcal{R}}(\theta(t)) \\ &= \theta(t) - \eta \frac{1}{S} \sum_{s \in \mathcal{S}} \nabla_{\theta(t)} F_s(\theta(t)). \end{aligned} \tag{3}$$

where $\nabla_{\theta(t)} \hat{\mathcal{R}}(\theta(t))$ is the estimated gradient of empirical risk on mini-batch and η is the learning rate.

Theoretical Analysis

In this section, we theoretically explore the impact of distributed data in federated learning on the optimal model size using a PAC-Bayes generalization bound for the stochastic algorithms. In particular, we establish the analytic solution of the optimal model size based on the proved bound, and compare the solutions between different scenarios to demonstrate several important insights.

Problem Setup

To align with the standard settings of Federated Learning (McMahan et al. 2017), we consider a distributed scenario consisting of n clients and a central server that connects all clients. Each client $i \in \{1, \dots, n\}$ possesses a local dataset $\mathcal{D}_i$, with the average dataset size denoted as $m = \frac{1}{n} \sum_{i=1}^{n} |\mathcal{D}_i|$. Thus, the total amount of data across all clients is nm. Suppose that training will be repeated for T rounds, following the classical FL algorithm FedAvg (McMahan et al. 2017), the training process at round $j \in \{1, \dots, T\}$ can be expressed as:

$$\begin{aligned} \bar{\theta}(j+1) &= \frac{1}{n} \sum_{i=1}^{n} \theta_i(j+1) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\bar{\theta}(j) - \eta \nabla_{\bar{\theta}(j)} \mathbb{E}_{\zeta_i \sim \mathcal{D}_i} F(\bar{\theta}(j); \zeta_i)). \end{aligned} \tag{4}$$

Eq.(4) demonstrates the formal update of FL in each round, with the first line denoting the model aggregation operation on the central server and the second line denoting the local training of the global model $\bar{\theta}(j)$ on clients. Besides, since the training optimization is performed through SGD, we also define the batch size for local training as $S_{Fed} = k_{Fed} m \in \{1, \dots, m\}$ where $\frac{1}{m} \leq k_{Fed} \leq 1$ and the number of local training epochs is t. Correspondingly, the baseline centralized scenario holds a dataset $\mathcal{D} = \bigcup_{i=1}^{n} \mathcal{D}_i$ of size nm, and the weights of the initial model in this scenario are the same as that in the federated scenario, i.e, $\{\theta(0) = \theta_i(0) | i \in n\}$. The training of θ follows the SGD optimization described in Eq.(3), denoted as:

$$\theta(j+1) = \theta(j) - \eta \nabla_{\theta(j)} \mathbb{E}_{\zeta \sim \mathcal{D}} F(\theta(j); \zeta), \tag{5}$$

and is iterated for $\frac{T}{n}$ rounds to match the total training compute, which we define following the previous scaling law studies (Kaplan et al. 2020; Muennighoff et al. 2023) as the total number of samples processed through training (i.e., dataset size times the number of training rounds). In each round, the model θ is trained using data from $\mathcal{D}$ for t epochs with the batch size $S_{Cen} = k_{Cen} nm \in \{1, \dots, nm\}$ where $\frac{1}{nm} \leq k_{Cen} \leq 1$. Furthermore, we have $k_{Fed} m \leq k_{Cen} nm$ due to more training data allocated to centralized settings in practice, leading to a generally larger batch size in use.

A Generalization Bound for Federated SGD

To prove a PAC-Bayes generalization bound for the stochastic algorithms under federated settings, we first present some common assumptions aligned with the previous research (Mandt, Hoffman, and Blei 2017; He, Liu, and Tao 2019).

Assumption 1 *Considering that the stochastic gradient $\hat{g}_s(\theta) = \nabla_{\theta(t)} \hat{\mathcal{R}}(\theta(t))$ is computed as the sum of S independent gradients uniformly sampled from the training dataset, we assume that the gradient noise is Gaussian with covariance $\frac{1}{S} C(\theta)$, so $\hat{g}_s(\theta)$ can be approximated as*

$$\hat{g}_s(\theta) \approx g(\theta) + \frac{1}{\sqrt{S}} \Delta g(\theta), \quad \Delta g(\theta) \sim \mathcal{N}(0, C(\theta)), \tag{6}$$

where $g(\theta)$ denotes the full gradient of the expected loss. We further assume that $C(\theta)$ remains approximately constant with respect to θ and can be factorized into:

$$C(\theta) \approx C = BB^{\top}, \tag{7}$$

where $C \in \mathbb{R}^{d \times d}$ is symmetric and (semi) positive-definite.

We justify Assumption 1 by the central limit theorem: when the training data size is much larger than the batch size, the aggregated mini-batch gradients make the SGD update noise approximately Gaussian with near-stationary covariance. Since deep neural networks are typically trained on large-scale datasets, this Gaussian assumption generally holds and is standard in OU-based SGD analyses (Ee 2017; Mandt, Hoffman, and Blei 2017). The constant matrix C is further justified when SGD iterates are confined to a small enough region around a local optimum of the loss, where the noise covariance varies little.

Assumption 2 *Assuming the loss function $F(\theta)$ is smooth, when the stationary distribution of the iterates is confined to a local region near a minimum θ^*, the loss gradient satisfies:*

$$\nabla F(\theta) \approx A(\theta - \theta^*), \tag{8}$$

where $A \in \mathbb{R}^{d \times d}$ is a constant (semi) positive-definite matrix representing the local Jacobian of the gradient field.

Assumption 2 is generally valid when SGD converges to a low-variance quasi-stationary distribution near a deep local minimum, where the gradient noise is small compared to the mean gradient. According to the fact that the exit time of a stochastic process is typically exponential in the barrier height between minima (Mandt, Hoffman, and Blei 2017), local optima remain stable even in the presence of noise, and SGD thus follows a relatively directed path toward the optimum. This behavior is also supported by empirical observations in a prior study (Li et al. 2018). Moreover, the assumption can be extended to general cases through a local translation. For example, translating around a stationary point θ^* with $\delta = \theta - \theta^*$ yields the Taylor form $\nabla F(\theta^* + \delta) = A\delta + O(\|\delta\|^2)$, assuming smoothness within a stable basin instead of global convexity.

Besides the above assumptions, we also need the formal definition of the PAC-Bayes upper bound to bound the generalization error. Following previous research (McAllester 1998, 1999), we have:

Lemma 1 *For any positive real $\delta \in (0,1)$, with probability at least $1-\delta$ over a sample of size N, we have the following inequality for the distribution of the output hypothesis Q and the prior P:*

$$R(Q) \leq \hat{R}(Q) + \sqrt{\frac{\mathcal{D}(Q||P) + \log(\frac{1}{\delta}) + \log(N) + 2}{2N-1}}, \tag{9}$$

where $\mathcal{D}(Q||P)$ is the KL divergence between the distributions Q and P and is defined as: $\mathcal{D}(Q||P) = \mathbb{E}_{\theta\sim Q}\log(\frac{Q(\theta)}{P(\theta)})$.

Based on the two assumptions and Lemma 1, we can prove the generalization bound for federated SGD below.

Theorem 1 *For any positive real $\delta \in (0,1)$, with probability at least $1-\delta$ over a distributed training data set with a total size nm across all clients, we have the following inequality for the distribution Q_{Fed} of the output hypothesis function of federated SGD:*

$$R(Q_{Fed}) - \hat{R}(Q_{Fed}) \leq \sqrt{\frac{H_1 + H_2 - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}}, \tag{10}$$

where

$$H_1 = -\log(\det(\Sigma_{Fed})), H_2 = \frac{T\eta}{2k_{Fed}m} tr(\bar{C}\bar{A}^{-1}), \tag{11}$$

d is the dimension of the parameter (the model size) and $tr(\bar{C}\bar{A}^{-1})$ is the trace of the product matrix $\bar{C}\bar{A}^{-1}$.

Apparently, it is hard to quantify the above generalization bound since the covariance matrix Σ_{Fed} for the stationary distribution is not available for the training data. In order to be able to estimate the optimal model size using the generalization bound, we introduce the following assumption and study a special case of the generalization bound as in other papers (He, Liu, and Tao 2019; Jastrzkebski et al. 2017).

Assumption 3 *We assume that A and Σ are symmetric matrices satisfying $A\Sigma = \Sigma A$.*

Assumption 3 implies that the local geometry around the global minimum and the stationary distribution is homogeneous across all dimensions of the parameter space, so A and Σ can be assumed to share eigenvectors. This common isotropic simplification, also adopted in prior papers (He, Liu, and Tao 2019; Jastrzkebski et al. 2017), makes trace and determinant terms analytically solvable without affecting the scaling order of the bound. When Assumption 3 also holds, we reformulate the property found in the proof of Theorem 1 and derive a new generalization bound as follows.

Theorem 2 *Under all the Assumptions of Theorem 1 and with Assumption 3, we have the following generalization bound for the stationary distribution of federated SGD:*

$$R(Q_{Fed}) - \hat{R}(Q_{Fed}) \leq \sqrt{\frac{H_{Fed} + H'_{Fed} - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}}, \tag{12}$$

where $H_{Fed} = d\log(\frac{2k_{Fed}m}{T\eta}) - \log(\det(\bar{C}\bar{A}^{-1}))$ and $H'_{Fed} = \frac{T\eta}{2k_{Fed}m} tr(\bar{C}\bar{A}^{-1})$.

Relationship between Two Optimal Model Sizes

In the same way, we can prove the generalization bound for centralized SGD with the same amount of training data.

Lemma 2 *Under all the assumptions of Theorem 2, we have the following generalization bound for the stationary distribution of centralized SGD trained on the same amount of training data:*

$$R(Q_{Cen}) - \hat{R}(Q_{Cen}) \leq \sqrt{\frac{H_{Cen} + H'_{Cen} - d + 2\log(\frac{1}{\delta}) + 2\log(nm) + 4}{4nm-2}}. \tag{13}$$

where $H_{Cen} = d\log(\frac{2k_{Cen}n^2m}{T\eta}) - \log(\det(CA^{-1}))$ and $H'_{Cen} = \frac{T\eta}{2k_{Cen}n^2m} tr(CA^{-1})$.

Generalization bounds provide an upper limit on an algorithm's generalization error, with smaller bounds indicating better generalization performance. A natural criterion for selecting the optimal model size d^* is the value of d that minimizes this bound. While convexity is not guaranteed in general, empirical studies on scaling laws (Kaplan et al. 2020; Hoffmann et al. 2022; Muennighoff et al. 2023) suggest an approximately convex relationship between model size and generalization performance. Based on this, we assume a locally convex regime and derive a closed-form approximation of d^* using first-order conditions.

Lemma 3 *When all the above assumptions hold, the optimal model size under the output hypothesis function of federated SGD has the following analytic solution:*

$$d^*_{Fed} = \frac{H_1 + H_2 + 8n\log(\frac{1}{\delta}) + 8n\log(nm) - \frac{4}{m} + 8n}{8n - \frac{2}{m} - 4n\log(\frac{2k_{Fed}m}{T\eta})}. \tag{14}$$

where $H_1 = -4n\log(\det(\bar{C}\bar{A}^{-1}))$ and $H_2 = (\frac{4nT\eta}{k_{Fed}m} - \frac{T\eta}{k_{Fed}m^2}) tr(\bar{C}\bar{A}^{-1})$.

On the other hand, the optimal model size for centralized SGD has the following analytic solution:

$$d_{Cen}^* = \frac{\hat{H}_1 + \hat{H}_2 + 8n\log(\frac{1}{\delta}) + 8n\log(nm) - \frac{4}{m} + 8n}{8n - \frac{2}{m} - 4n\log(\frac{2k_{Cen}n^2m}{T\eta})}. \tag{15}$$

where $\hat{H}_1 = -4n\log(\det(CA^{-1}))$ *and* $\hat{H}_2 = (\frac{4T\eta}{k_{Cen}nm} - \frac{T\eta}{k_{Cen}n^2m^2})tr(CA^{-1})$.

The comparison between d_{Fed}^* and d_{Cen}^* reveals their relationship. However, we cannot quantify this relationship without knowing the connection between the average local distribution and the global distribution (i.e., $\text{tr}(\bar{C}\bar{A}^{-1})$ vs $\text{tr}(CA^{-1})$)). Hence, we introduce another assumption.

Assumption 4 *Under the fair comparison condition that the same training dataset is used for both training scenarios, we assume that the local data distributions* $\mathcal{D}_1, \dots, \mathcal{D}_n$ *across* n *clients of size* m *form a heterogeneous partition of the global dataset* $\mathcal{D}$ *of size* $D = nm$*. Hence, we have the following approximate relationships:*

$$\bar{A} \approx A + \Delta_A, \quad \bar{C} \approx \frac{1}{n^\gamma}(C + \Delta_C), \tag{16}$$

where $\gamma > 1$*, and* Δ_A*,* Δ_C *are deviation terms introduced by data heterogeneity. These deviations are assumed to be bounded in norm:*

$$\|\Delta_A\| \le \epsilon_A, \quad \|\Delta_C\| \le \epsilon_C, \tag{17}$$

where ϵ_A, ϵ_C *grow with the non-i.i.d degree across clients.*

Assumption 4 reflects realistic FL scenarios where client data are drawn from heterogeneous (non-i.i.d.) distributions. It can be justified by the central limit theorem when both the average local data size m and the global dataset D are large. While the centralized matrices A and C describe curvature and gradient noise under the full dataset, their decentralized averages $\bar{A}$ and $\bar{C}$ deviate under non-i.i.d. sampling. Following the bounded-variance formulation in FL theory (Karimireddy et al. 2020), we include deviation terms Δ_A and Δ_C whose magnitudes quantify heterogeneity, and introduce the scaling exponent $\gamma > 1$ to capture how such heterogeneity amplifies the gap between local and global covariances while keeping the analysis analytically tractable. Under this assumption, the relationship between the two optimal model sizes is shown below.

Theorem 3 *When all the above assumptions hold, by comparing the optimal model size between the federated and centralized scenarios, we find that:*

$$\lim_{T\to\infty} d_{Fed}^* = \frac{\rho}{n^{\gamma-1}} d_{Cen}^*, \tag{18}$$

where $\rho = \frac{S_{Cen}\left(tr(CA^{-1}) + tr(\Delta_1)\right)}{S_{Fed}tr(CA^{-1})} > 0$ *and* $\Delta_1 = (CA^{-1}\Delta_A + \Delta_C(I + A^{-1}\Delta_A))A^{-1}$.

Remark 1 *Since we have* $\gamma > 1$*, Theorem 3 shows* $d_{Fed}^* < d_{Cen}^*$ *and suggests the* ***first theoretical insight:***

- *When transferring the training of large-scale models from centralized to federated scenarios with the same training compute, the optimal model size should be decreased, and the reduction ratio has a negative power law relationship with the number of clients.*

Theoretical Evidence for Generalization Gap

The above theoretical proofs demonstrate the generalization bound and the optimal model size under this bound for each training. In this subsection, we show that these proofs can also serve as an important theoretical basis for an empirical finding observed in many previous works. Specifically, models trained with distributed data are generally found to be inferior to the models trained with centralized data in performance (Lian et al. 2017; Sun, Li, and Wang 2022). According to the respective generalization bound and optimal model size, we derive the following theorem.

Theorem 4 *When all the above assumptions hold, we find the following inequality between the optimal generalization error of federated SGD and centralized SGD using the same training compute:*

$$\lim_{T\to\infty} (\mathcal{G}_{Fed}^* - \mathcal{G}_{Cen}^*) > 0 \tag{19}$$

when the number of clients n *satisfies the property:* $n > \sqrt[\gamma-1]{\rho}$*. Here,* $\mathcal{G}^*$ *is the optimal generalization error computed with the optimal model size* d^**.*

Remark 2 *The condition in Theorem 4 basically holds in practice, considering that realistic federated scenarios generally scale to a sufficiently large number of clients (Kairouz et al. 2021) (e.g., phones with user data, edge sensors, etc.) Also, it is expected that the value of* ρ *would not be very large. In the ideal case where client data is i.i.d. and both training uses identical batch size, we have* $\rho = 1$*. Since* $n \ge 2$ *holds for any federated scenarios, the inequality will be trivially satisfied. Based on this result, we summarize our* ***second theoretical insight:***

- *If a federated scenario with a large number of clients is not allocated more data than the centralized scenario, data decentralization will lead to a definite gap between the optimal generalization performance achieved through FL and that under centralized settings, which underscores the challenges of FL.*

Estimating Optimal Model Size by the Average Training Compute Between Clients

Previous analyses have studied the optimal model size for federated SGD training using all local data from the clients. Notably, the total training compute for the federated SGD training is equal to the sum of the training compute allocated to each client. Therefore, we are also interested in the optimal model size at the local level and how it relates to the above results. By a similar proof, we derive the analytic solution of the optimal model size at the client level as follows.

Lemma 4 *When all the above assumptions hold, the optimal model size at the client level* d_i^* *has the following analytic solution:*

$$d_i^* = \frac{H_1^{(i)} + H_2^{(i)} + 8\log(m) - \frac{4}{m} + 8}{8 - \frac{2}{m} - 4\log(\frac{2k_im}{T\eta})}, \tag{20}$$

where $H_1^{(i)} = -4\log(\det(C_iA_i^{-1}))$ *and* $H_2^{(i)} = (\frac{4T\eta}{k_im} - \frac{T\eta}{k_im^2}) + 8\log(\frac{1}{\delta}) - 4\log(\det(C_iA_i^{-1}))$.

Then, considering that the local data on clients is heterogeneous, we use $\xi_i^C = C_i - \bar{C}$ and $\xi_i^A = A_i - \bar{A}$ to denote client variance in non-IID settings. Comparing d_i^* with the optimal model size d_{Fed}^* in FL across n clients shows their relationship as follows.

Theorem 5 *When all the above assumptions hold, considering the same batch size $\{k_{Fed}m = k_i m | i \in n\}$, the following relation holds between the optimal model size d_i^* on a single client and the optimal model size d_{Fed}^* of FL across n clients:*

$$\lim_{T\to\infty} d_{Fed}^* = \frac{\kappa}{n}\sum_{i=1}^{n} d_i^*, \tag{21}$$

where $\kappa = \frac{\left(4m-\frac{1}{n}\right)tr(\bar{C}\bar{A}^{-1})}{(4m-1)\left(tr(\bar{C}\bar{A}^{-1})+tr(\bar{\xi})\right)} > 0$ *and* $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}((\bar{C}\bar{A}^{-1}\xi_i^A + \xi_i^C(I + \bar{A}^{-1}\xi_i^A))\bar{A}^{-1})$.

Remark 3 *Since the existing scaling law suggests that the optimal model size relates to the data size (Kaplan et al. 2020), it is intuitive that the optimal model size in FL would be decided by the total data size across clients (i.e., $d_{Fed}^* \approx \sum_{i=1}^{n} d_i^*$). However, Theorem 5 demonstrates that this thought is incorrect. Eq.(21) implies $d_{Fed}^* \approx \frac{1}{n}\sum_{i=1}^{n} d_i^*$, as $\frac{4m-\frac{1}{n}}{4m-1} \approx 1$ and the average bias term $tr(\bar{\xi})$ is much smaller than $tr(\bar{C}\bar{A}^{-1})$. This result highlights our* ***third theoretical insight:***

- *The optimal model size in FL is primarily determined by the average training compute per client, rather than the total compute across all clients or the number of clients.*

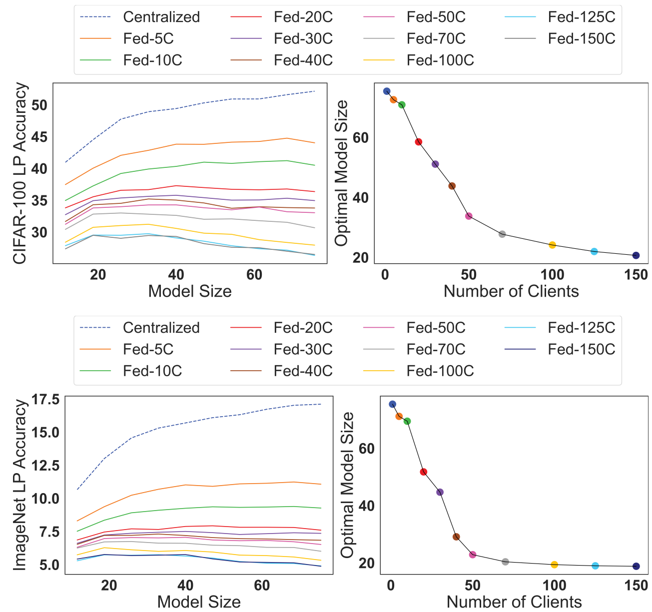

Figure 1: Impact of distributed data on the optimal model size of ViT. (Top / (a)) CIFAR-100 Results. (Bottom / (b)) ImageNet Results. (Left) Curves of LP accuracy (%) versus model size. Different lines represent FL scenarios with a different number of clients. (Right) Curve of optimal model size versus the number of clients. Here, the centralized setting corresponds to the case $n = 1$.

Empirical Validation

Experiment Setup

We conduct multiple experiments based on a popular model architecture, Vision Transformer (ViT) (Dosovitskiy et al. 2021). This architecture represents a dominant type of model in deep learning: Transformers (Vaswani et al. 2017) relying on the attention mechanism, which is frequently used for building large-scale models. Specifically, we build 10 different sizes of ViTs with parameters ranging from 11.62 to 75.41 million. These models are pre-trained on the Mini-Imagenet dataset (Vinyals et al. 2016), which contains 60,000 images extracted from the ImageNet dataset (Deng et al. 2009). We adopt the Masked Auto-encoder (He et al. 2022) (MAE) approach to pre-train ViTs. To evaluate the effectiveness of pre-training, we conduct linear probing tests, which freeze the pre-trained weights in the backbone and only fine-tune the head layer (He et al. 2016). The linear probing accuracies of these models on two standard datasets (CIFAR-100 (Krizhevsky and Hinton 2009) and ImageNet (Deng et al. 2009)) are collected for analysis. We select the size of the model with the highest linear probing (LP) accuracy as the optimal model size.

For all experiments, we strictly follow the problem setup defined in the theoretical analysis. All training resources are kept the same between the centralized and federated scenarios, including the model, total training compute, and dataset. To simulate federated scenarios with n clients and non-IID client data, we divide the training dataset into n partitions by sampling the class priors of the Dirichlet distribution (Hsu, Qi, and Brown 2019). A more heterogeneous division can be made by specifying a smaller Dirichlet parameter α during sampling. We use $\alpha = 0.1$ by default. Our codes for experiments were implemented using the PyTorch framework and executed on a server with 8 NVIDIA® RTX A5000 GPUs.

Empirical Results

Evidence for the First Insight. We investigate the impact of distributed data on the optimal model size by training models with the same training compute in both the centralized scenario and federated scenarios with different numbers of clients. Figure 1(a, Left) shows the linear probing accuracies of ViTs with different sizes on CIFAR-100 in each scenario. Based on the highest accuracy, we find the optimal model size for each scenario and plot them in Figure 1(a, Right). The results clearly show a negative power-law relationship between the optimal model size and the number of clients, validating Theorem 3. Besides, we have also collected the linear probing results of ViTs on the ImageNet dataset (Deng et al. 2009), which has around 1.2 million images. We use 10% of the training samples for linear probing and still observe similar empirical results, as shown in Figure 1(b). To further validate the size relationship, we fit the log–log form of Theorem 3 (i.e., $\log d_{Fed}^* = (1-\gamma)\log n + \log\rho + \log d_{Cen}^*$) by the measured optimal sizes in FL. This gives slopes in $[-0.28, -0.48]$ ($\gamma \in [1.3, 1.5]$, $\rho \in [2.0, 2.4]$) with coefficient of determination $R^2 > 0.88$, confirming a strong

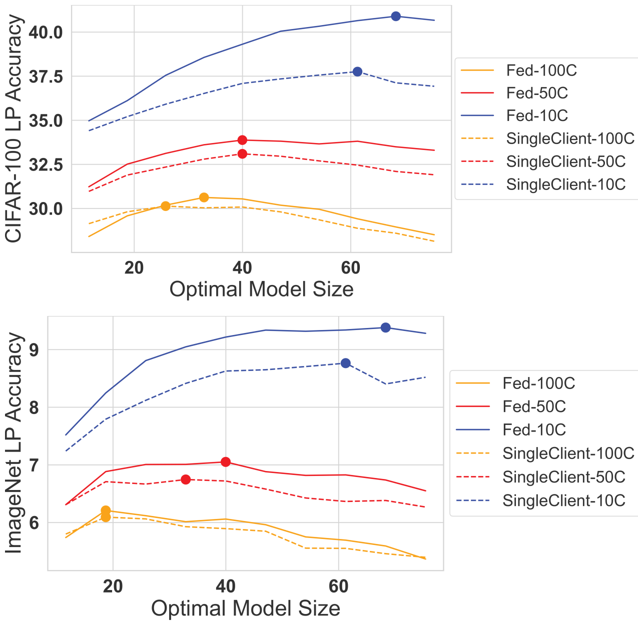

Figure 2: Comparison between the optimal model size across all clients and for a single client. The dots represent the highest accuracy of each line. (Top / (a)) CIFAR-100 Results. (Bottom / (b)) ImageNet Results.

negative power law in n.

Evidence for the Second Insight. Figure 1(Left) also shows the impact of the number of clients on the linear probing accuracy of models. We note that when the federated scenario does not have an advantage on the size of training data (i.e., the total training data across clients equals the training data used in centralized training), the models trained by federated learning will be inferior to the models trained by centralized learning in terms of the generalization performance even if the number of clients is very small (a good example here is $n = 5$). Moreover, the gap between the two training regimes broadens as the number of clients increases. These findings are consistent with our theoretical results in Theorem 4.

Evidence for the Third Insight. FL indirectly uses all training data from clients to train a global model by an iterative process of having multiple models trained on different clients using their local data and aggregating the parameters of these models on the server. In Figure 2, we train models using only local data from a single client and compute the average of n sets of linear-probing accuracies from n clients. These results are then compared with those of FL. We observe that the optimal model size is actually very close between the two training cases, which matches our third insight shown by Theorem 5. Thus, if the optimal model size in a centralized scenario with the same amount of training data is not known in advance, the optimal model size in a federated scenario can also be estimated based on the average training compute allocated to each client.

Applicability Study. Beyond ViTs, we evaluate whether our theoretical insights hold for convolutional models. Figure 3 demonstrates that adapting our theoretical findings to

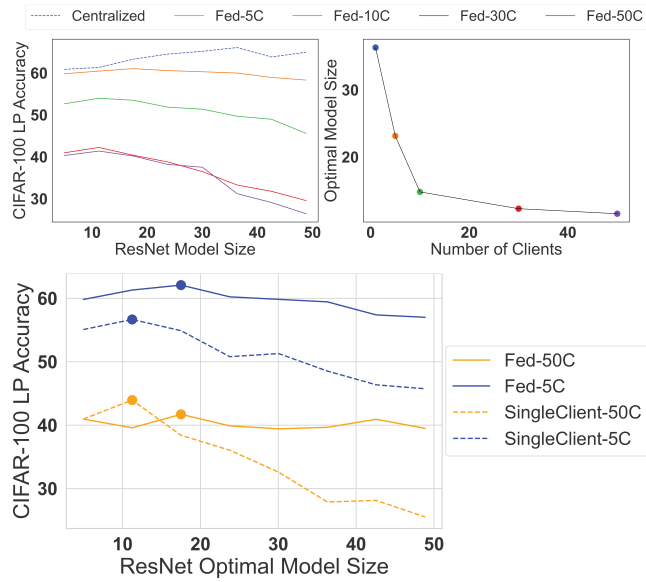

Figure 3: Applicability Analysis on ResNets. (Top / (a)) Impact of distributed data on the optimal model size. (Bottom / (b)) Comparison between the optimal model size across all clients and for a single client.

ResNets (He et al. 2016) exhibits similar behavior on their optimal model sizes, reinforcing that the derived insights are not tied to a specific architecture and may thus serve as general guidelines for size selection in distributed training.

Conclusion

This work investigates the scaling behavior of large-scale models in federated learning, with a focus on how distributed data affects the estimation of the optimal model size. We derive a PAC-Bayes generalization bound for federated SGD and analyze the global optimal model size under this bound. From this, we obtain three main insights. First, data decentralization reduces the optimal model size, following an approximate negative power law with respect to the number of clients. Second, moving large-scale training to federated settings inevitably lowers the achievable generalization performance. Third, the optimal model size should be estimated by the average training compute per client rather than the total compute or network size. Various experiments on transformer and convolutional models across multiple datasets confirm these findings. We expect our results to offer practical guidance for deploying large-scale models in distributed environments.

Acknowledgments

This work is partly supported by the Australian Research Council Linkage Project (Grant No. LP220200893) and the National Key R&D Program of China (Grant No. 2021YFA1000402).

References

AbdulRahman, S.; Tout, H.; Ould-Slimane, H.; Mourad, A.; Talhi, C.; and Guizani, M. 2021. A Survey on Federated Learning: The Journey from Centralized to Distributed On-Site Learning and Beyond. *IEEE Internet of Things Journal*, 8(7): 5476–5497.

Achituve, I.; Shamsian, A.; Navon, A.; Chechik, G.; and Fetaya, E. 2021. Personalized Federated Learning with Gaussian Processes. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, 8392–8406.

Boroujeni, M. G.; Krause, A.; and Ferrari Trecate, G. 2024. Personalized Federated Learning of Probabilistic Models: A PAC-Bayesian Approach. arXiv:2401.08351.

Bottou, L. 1998. Online Learning and Stochastic Approximations. In Saad, D., ed., *Online Learning in Neural Networks*, 9–42. Cambridge University Press.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; and et al. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 1877–1901.

Chen, C.; Feng, X.; Zhou, J.; Yin, J.; and Zheng, X. 2023. Federated Large Language Model: A Position Paper. arXiv:2307.08925.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; and et al. 2021. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ee, W. 2017. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5(1): 1–11.

Gadre, S. Y.; Smyrnis, G.; Shankar, V.; Gururangan, S.; Wortsman, M.; Shao, R.; Mercat, J.; Fang, A.; Li, J.; Keh, S.; et al. 2024. Language Models Scale Reliably with Over-Training and on Downstream Tasks. arXiv:2403.08540.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

He, F.; Liu, T.; and Tao, D. 2019. Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollar, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hilmkil, A.; Callh, S.; Barbieri, M.; Sutfeld, L. R.; Zec, E. L.; and Mogren, O. 2021. Scaling Federated Learning for Fine-Tuning of Large Language Models. In *Proceedings of the International Conference on Applications of Natural Language to Information Systems (NLDB)*, 15–23. Springer.

Hinton, G. E.; and Salakhutdinov, R. R. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786): 504–507.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; and et al. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2019. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. arXiv:1909.06335.

Jastrzkebski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2017. Three Factors Influencing Minima in SGD. arXiv:1711.04623.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D.; and et al. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S. U.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, 5132–5143. PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009, University of Toronto.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the Loss Landscape of Neural Nets. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 31.

Lian, X.; Zhang, C.; Zhang, H.; Hsieh, C.-J.; Zhang, W.; and Liu, J. 2017. Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

London, B. 2017. A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30.

Mandt, S.; Hoffman, M. D.; and Blei, D. M. 2017. Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*, 18(134): 1–35.

McAllester, D. A. 1998. Some PAC-Bayesian Theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, 230–234.

McAllester, D. A. 1999. PAC-Bayesian Model Averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, 164–170.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Aguera y Arcas, B. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282. PMLR.

Mou, W.; Wang, L.; Zhai, X.; and Zheng, K. 2018. Generalization Bounds of SGLD for Non-Convex Learning: Two Theoretical Viewpoints. In *Proceedings of the Conference on Learning Theory (COLT)*, 605–638. PMLR.

Muennighoff, N.; Rush, A.; Barak, B.; Le Scao, T.; Tazi, N.; Piktus, A.; Pyysalo, S.; Wolf, T.; and Raffel, C. A. 2023. Scaling Data-Constrained Language Models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, volume 36.

Pensia, A.; Jog, V.; and Loh, P.-L. 2018. Generalization Error Bounds for Noisy, Iterative Algorithms. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 546–550. IEEE.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; and et al. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. arXiv:2112.11446.

Ro, J. H.; Breiner, T.; McConnaughey, L.; Chen, M.; Suresh, A. T.; Kumar, S.; and Mathews, R. 2022. Scaling Language Model Size in Cross-Device Federated Learning. arXiv:2204.09715.

Sefidgaran, M.; Chor, R.; Zaidi, A.; and Wan, Y. 2024. Lessons from Generalization Error Analysis of Federated Learning: You May Communicate Less Often! In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR.

Shen, T.; Zhu, D.; Zhao, Z.; Li, Z.; Wu, C.; and Wu, F. 2025. Will LLMs Scaling Hit the Wall? Breaking Barriers via Distributed Resources on Massive Edge Devices. arXiv:2503.08223.

Sun, T.; Li, D.; and Wang, B. 2022. Adaptive Random Walk Gradient Descent for Decentralized Optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 20790–20809. PMLR.

Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the Importance of Initialization and Momentum in Deep Learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 1139–1147. PMLR.

Tang, Z.; Shi, S.; Li, B.; and Chu, X. 2023. GossipFL: A Decentralized Federated Learning Framework with Sparsified and Adaptive Communication. *IEEE Transactions on Parallel and Distributed Systems*, 34(3): 909–922.

Tay, Y.; Dehghani, M.; Rao, J.; Fedus, W.; Abnar, S.; Chung, H. W.; Narang, S.; Yogatama, D.; Vaswani, A.; and Metzler, D. 2021. Scale Efficiently: Insights from Pre-Training and Fine-Tuning Transformers. arXiv:2109.10686.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; and et al. 2022. LaMDA: Language Models for Dialog Applications. arXiv:2201.08239.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30.

Vedadi, E.; Dillon, J. V.; Mansfield, P. A.; Singhal, K.; Afkanpour, A.; and Morningstar, W. R. 2024. Federated Variational Inference: Towards Improved Personalization and Generalization. In *Proceedings of the AAAI Symposium Series*, volume 3, 323–327. AAAI Press.

Vinyals, O.; Blundell, C.; Lillicrap, T.; and Wierstra, D. 2016. Matching Networks for One-Shot Learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 29.

Wu, X.; Liang, Z.; and Wang, J. 2020. FedMed: A Federated Learning Framework for Language Modeling. *Sensors*, 20(14): 4048.

Xu, M.; Cai, D.; Wu, Y.; Li, X.; and Wang, S. 2024. FwdLLM: Efficient Federated Finetuning of Large Language Models with Perturbed Inferences. In *Proceedings of the USENIX Annual Technical Conference (USENIX ATC)*, 579–596. USENIX Association.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12104–12113.

Zhao, Z.; Liu, Y.; Ding, W.; and Zhang, X.-P. 2024. Federated PAC-Bayesian Learning on Non-IID Data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5945–5949. IEEE.

Zhu, M.; Shen, L.; Du, B.; and Tao, D. 2024. Stability and Generalization of the Decentralized Stochastic Gradient Descent Ascent Algorithm. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 36.