Data Quality Control in Federated Instruction-tuning of Large Language Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

By leveraging massive distributed data, federated learning (FL) enables collaborative instruction tuning of large language models (LLMs) in a privacy-preserving way. While FL can effectively extend the data quantity, data quality, despite its significance, is under-explored in current literature of FL of LLMs. In response, we propose a new framework of federated instruction tuning of LLMs with data quality control (FedDQC), which measures data quality to facilitate the subsequent processes of filtering and hierarchical training. Specifically, we firstly propose an efficient data quality evaluation metric, which measures per-sample instruction-response alignment (IRA) at each client side with a single-shot inference. Based on this, samples with relatively low IRA is potentially noisy data, therefore are filtered to mitigate their negative impacts. To further utilize this IRA value, we propose a quality-aware hierarchical training paradigm, where the LLM is progressively fine-tuned from high-IRA to low-IRA data, mirroring the easy-to-hard learning process of humans. We conduct extensive experiments on 4 domain-specific datasets and a general dataset, and compare our method with baselines that are adapted from centralized learning. Results show that our method consistently and significantly improves the performance of LLMs that are trained on mix-quality data via FL.

1 Introduction

For large language models (LLMs) training [1, 2, 3, 4], both the quantity and quality of the training data significantly impact their performance [5, 6]. The scaling law suggests that more training data can lead to more powerful LLMs [7]. However, in specific domains such as healthcare [8] and finance [9], privacy concerns [10] prevent the aggregation of large-scale datasets, making it challenging to expand the dataset scale. Federated Learning (FL) [11], an emerging distributed training approach, preserves privacy by allowing multiple clients to train a unified model collaboratively without sharing their data. This enables dataset scaling while ensuring data privacy [12, 13].

While FL addresses the data quantity issue by incorporating more local clients, it may introduce more data quality issues [14]. In FL, training data for each client are collected from various sources locally, making it difficult to detect low-quality data or errors in local datasets. Such vulnerabilities may adversely affect general model training. Although numerous methods [15, 16, 17, 18, 19] are proposed for data quality control in LLM training, their designs typically require access to the entire training data, making them impractical for FL scenarios. Therefore, in this work, we aim to bridge this gap and address the under-explored issue of federated data quality control in instruction-tuning LLM tasks.

Existing data quality control methods for instruction tuning focused on designing data quality evaluation metrics [20]. These metrics aim to quantify the quality of instruction-response pairs. However, these metrics are not suitable for federated settings for two main reasons. First, adapting these metrics to FL might compromise privacy and computational efficiency. For instance, [21, 16] using external models for evaluation can breach privacy as it involves sending private data out for assessment. Moreover, methods [22, 23] using the influence function [24] require extensive computation, which is impractical in FL where resources are often limited. Second, these metrics are disconnected from the training process. They focus on identifying the most informative yet challenging data points, which often include noisy data mistakenly identified as high-quality [25, 26, 17]. This can destabilize the training process.

To address these two issues in controlling data quality in federated instruction-tuning, we propose a new framework Federated Data Quality Control (FedDQC). The key idea is to directly measure instruction-response relativeness as an aspect of quality evaluation and integrate this measurement with easy-to-hard hierarchical training. Our framework has two main innovative designs: 1) an alignment-based data quality evaluation metric that is computation-efficient without violating privacy. Firstly, motivated by mutual information [27] we proposed the Instruction-Response Alignment (IRA), defining the data quality from the aspect of the relativeness between the instruction and response. In this aspect, the high-scored samples are instruction-response, highly related, and easy to learn. Before the start of training, each client uses the initial global model to evaluate per-sample data quality and select the high-scored samples with the global threshold. This guarantees a globally unified data quality evaluation and data selection standard. 2) A quality-aware hierarchical training that follows an easy-to-hard paradigm. It forces the model to prioritize learning from these highly related samples and then move on to less related, complex data. By starting training with highly aligned data, it helps the model capture the intrinsic features of the data when dealing with varying quality, preventing overfitting to specific errors or noise. This method makes learning more efficient and ensures a more robust model. In addition, by setting the participants in each round to have similar data difficulty, this hierarchical training enables clients to have similar data patterns, thus accelerating convergence.

Our experiments demonstrate that FedDQC not only outperforms all baseline models in both IID (independent and identically distributed) and non-IID settings on four domain-specific datasets but also shows effectiveness on the general dataset, Alpaca-GPT4 [28]. As for computation, we show that the scoring metric IAR consumes only 1% training time for data quality evaluation, making it computation-efficient and scalable for larger datasets.

The contributions of this work are three folds:

- We are the first attempt to tackle the practical issue of federated data quality control in LLM instruciton-tuning.
- We propose an FL data quality control framework FedDQC, which integrates alignment-based quality assessment with quality-aware hierarchical training to enhance efficient and robust instruction-tuning in federated scenarios without violating privacy and computation constraints.
- We experiment on four datasets from specific domains and a general domain dataset, showing the effectiveness of FL data quality control in both IID and NIID scenarios.

2 Related work

2.1 Federated Learning

Federated Learning [29] has emerged as a powerful method for privacy-preserving collaborative training, allowing multiple clients to jointly train a global model without sharing raw data, coordinated by a central server. Existing research on data quality in FL primarily focused on the classification tasks, with noisy label issues. [30] We classify related data quality control works from three levels: client, model and sample level. At the client level, efforts have concentrated on identifying malicious clients [31, 32] through feature [33] or model weight clustering [34]. While at the sample level, studies have typically focused on label correction strategies [35] or confidence-based sample reweighting [36]. At the model level, approaches like distillation [37] or modifying the loss function [38] aimed to increase robustness against noisy labels. However, these methods do not effectively address the unique challenges of federated LLM training, the generation task. This highlights the gap in current approaches and underscores the need for specialized solutions tailored to generative tasks in FL.

2.2 Data quality control

Data quality control is complex and a throughout problem in machine learning [39]. To solve the task for this work, we split the related work into two lines: the traditional data attribution with its adaptation to LLM setting, and current data selection work for LLM.

Data attribution Traditional data attribution methods, used to explain model predictions by identifying influential training examples, are generally categorized into retraining-based and gradient-based techniques. [40] Retraining-based approaches, such as leave-one-out [41], Shapley value [42], and Datamodels [43], estimate the effect of data points by repeatedly retraining the model on different subsets of data. These data attribution approaches are post-hoc and computation costly, making them unsuitable for LLM setting. Gradient-based approaches, like represented point selection [44], TracIn [45], and influence functions [24], estimate training data's impact through parameter sensitivity. Recent studies have developed more efficient adaptations of this gradient-based method for generative tasks [46] and LLM settings, streamlining data selection processes such as pre-training [47] and instruction-tuning in transfer learning scenarios [23]. Despite these advancements in reducing computational complexity through approximations, computing these methods for LLM data selection is still costly due to the increasing size of large model and data volumes.

Data selection for LLMs Current data selection works for LLM instruction-tuning are heuristic and aimed at core set selection. They either depend on a powerful external model for scoring or require iterative training or selection. External model-based scoring techniques, such as AlpaGasus [21], DEITA [16] and INSTAG [48] prompt ChatGPT [1] for various dimension of data quality scoring. While effective, these methods are costly and compromise privacy by requiring direct data sharing. This is particularly problematic in privacy-sensitive settings. Other methods that comply with privacy constraints still require large computation and are not well-suited for local dataset management essential in FL environments. For instance, IFD [25] and MoDS [26] require a computationally intensive initial training stage that may involve low-quality data. Similarly, InstructionMining [17] despite utilizing innovative statistical regression to fit quality influence factors with performance, is dataset-specific and requires retraining. Additionally, approaches like SelectIT [49] and NUGGETS [50] utilize in-context learning but highly depend on the predefined task set, which is sometimes applicable for FL. These challenges underscore the need for a new, locally implementable, efficient scoring method that preserves privacy and reduces computational overhead.

3 Problem formulation

3.1 Background of Federated Learning

In federated instruction tuning, each client holds an instruction-tuning dataset, where each sample is a pair of an instruction $(question, answer)$. Suppose there are total N clients, where the n-th client hold private dataset $\mathcal{D}_n = \{(q^i, a^i) | i = 1, 2, \ldots, |\mathcal{D}_n|\}$ and a local model θ_n^r, where q^i and a^i denote the i-th instruction and answer, r denote the round indices of training. Mathematically, the global objective of federated learning is $\min_\theta L(\theta) = \sum_{n=1}^{N} w_n L_n(\theta)$, where $w_n = \frac{|\mathcal{D}_n|}{\sum_{i=1}^{N} |\mathcal{D}_i|}$ and $L_n(\theta)$ are the dataset relative size and local objective of client n, respectively. The instruction-tuning training loss for the i-th sample is formulated as $L((a^i, q^i), \theta) = -\sum_{j=1}^{l_i} \log p(a_j^i | q_i \oplus a_{<j}^i; \theta)$, where $\oplus$ is the concatenation operator, l_i is the token length of output a^i and $a_{<j}^i$ denotes the tokens before index j. In the basic FL, FedAvg each training round r proceeds as follows: 1) Sever broadcasts the global model θ^r to clients; 2) Each client n performs local model training using t SGD steps to obtain a trained model denoted by $\theta^{r,t}$; 3) Clients upload the locally trained models $\theta^{r,t}$ to the server and the server updates the global model based on the aggregated local model: $\theta^{r+1} = \sum_{n=1}^{N} w_n \theta_n^{r,t}$.

3.2 Data quality control in FL

The data quality control problem is equivalent to reweighting training samples to achieve the best performance on the test set. In the federated setting, we could formulate the federated data quality control problem as follows:

$$\min_{\pi(\cdot)} \mathbb{E}_{(q_T^i, a_T^i) \in \mathcal{D}_T} L((q_T^i, a_T^i), \theta^*) \quad \text{s.t. } \theta^* = \min_\theta \sum_{n=1}^{N} w_n \mathbb{E}_{(q^i, a^i) \in \mathcal{D}_n} \pi((q^i, a^i)) L((q^i, a^i), \theta)$$

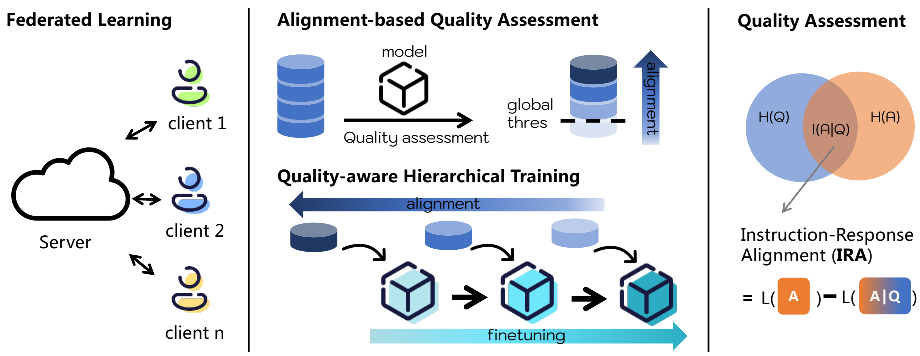


Figure 1: Overview of FedDQC, which consists of two components at the client side. (1) Alignment-based quality assessment: measuring the data quality from the instruction-response alignment perspective. This calculation is analogous to estimating mutual information between instruction and response, illustrated in the right figure; (2) Quality-ware hierarchical training: progressively fine-tuned from high-IRA to low-IRA data, mirroring the easy-to-hard learning process.

where $\pi(\cdot) : \mathcal{D} \rightarrow \mathcal{W}$ is the data quality control function with $\mathcal{W}$ represents the weight set, $\mathcal{D}_T$ and (q_T^i, a_T^i) are the test set and test sample i. The data control quality function could be written as a composition of two functions, $\pi((q^i, a^i)) = g(f((q^i, a^i)))$, where $f : \mathcal{D} \rightarrow \mathcal{S}$ is the scoring function where $\mathcal{D}$ denotes the dataset, $\mathcal{S}$ denotes the score set, and the reweighting function $g : \mathcal{D} \times \mathcal{S} \rightarrow \mathcal{W}$. The scoring function maps each data point with a scalar indicating the data quality and the reweighting function assigns weight to the data sample according to the quality score.

4 Methodology

To address data quality control in FL within computational and privacy constraints, we propose a two-stage FL data quality control framework, FedDQC. Section 4.1 will first give an overview of this framework, Section 4.2 and Section 4.3 will discuss the two components in detail. Lastly, computation, privacy and communication of FedDQC and comparisons with current methods will be discussed in Section 4.4.

4.1 Overview

To control the data quality for training, two steps need to be conducted: data selection based on data quality assessment and high-quality data training. Since in FL, data is preserved at the client side, only the client could assess their data quality and select its data based on the data quality score. In our FedDQC framework, data manipulations are mainly on the client side including the data quality measurement and local data training. The key idea of this framework is to integrate data quality assessment with the training process, which consists of two components the alignment-based data quality assessment and the quality-aware hierarchical training. These components are detailed in Algorithm 1 and illustrated in Figure 1.

The entire pipeline is described as follows. Initially, the server distributes the initial global model θ^0 to each client. Before training, clients assess their data quality using the global model and the data quality evaluation metric f to achieve consistent global data quality measurement: $s_i = f((q^i, a^i), \theta^0)$ for $(q^i, a^i) \in \mathcal{D}_n$. After assigning each data with a quality score s_i, clients select their local data based on a global threshold λ. The selected data are later sorted in descending order and split into K separate hierarchies, $\mathcal{H}_{n1}, \ldots, \mathcal{H}_{nk}$. In the training stage, FedDQC adheres to the FedAvg [11] local update and aggregation paradigm but changes the local dataset training sequence. On the client side, rather than random batching from the whole dataset, each client locally trains their models from the hierarchy $\mathcal{H}_{n1}$ with the highest score of data to the hierarchy $\mathcal{H}_{nK}$ with the lowest score. On the server side, during each round's aggregation, every client updates their model, which is trained on the same level of hierarchy. This guarantees synchronized, easy-to-hard training across all clients and a more consistent aggregation.

4.2 Alignment-based Quality Assessment

As an attempt to control data quality in federated learning, FedDQC needs to assess the data quality under the constraints of privacy and limited computation. We propose a novel data quality evaluation metric, the Instruction-Response Alignment (IRA), which leverages the concept of mutual information [27] to assess the alignment between instructional prompts and responses within local dataset. Specifically, IRA utilizes the initial global model to calculate the difference in loss between unconditioned responses and responses conditioned on their corresponding instructions. The following equation redefines the scoring function f to f_{IRA}:

$$f_{IRA}((q^i, a^i) \in \mathcal{D}, \theta) = L(a^i; \theta) - L((a^i, q^i); \theta)$$

where $L(a^i;\theta) = -\sum_{j=1}^{l_i} \log p(a^i_j | a^i_{<j}; \theta)$ calculates the loss of generating response a^i without given the instruction q^i, $L((a^i, q^i); \theta)$ is the loss given instruction q^i, which is defined in Section 3.1. $\mathcal{D}$ is dataset and θ represents model parameter for data quality evaluation.

This metric relates to the instruction, response, and initial training model to integrate data quality value with learning difficulty. High-score samples usually demonstrate a strong instruction-response relativeness from the perspective of the training model, indicating its learning is easier for the model. This quality evaluation does not compromise privacy and is more computation-efficient than the existing data evaluation metric. Later, based on the global threshold λ, clients select local data for training with the threshold above λ.

4.3 Quality-aware Hierarchical Training

After data selection, the pivotal next step is to start training. In this stage, we propose quality-aware hierarchical training, inspired by the philosophy of curriculum learning (CL) [51], where models learn progressively from easier to harder data, similar to human curricula.

As previously discussed, IRA evaluates instruction-response relativeness and indicates the relative difficulty of model training. Utilizing this metric, we could easily split training data into several hierarchies. The LLM begins with learning basic, instruction-response highly relevant problems, then progresses to applying the learned instruction-following ability to more generalized problems and gradually advances to more complex problems. Moreover, compared to the precisely defined learning scheduler [52], in CL, which precisely decides the sequence of data subsets throughout training, our hierarchical training is more coarse-grained. The number of hierarchies is typically set between 2 to 5, simplifying implementation across various clients.

We summarize the advantages of this method in three ways. 1) This method ensures that the model first establishes a strong foundational understanding, enhancing overall learning effectiveness and robustness; 2) It also ensures a consistent quality of data in each training round, which helps prevent the divergence of the aggregated model; 3) By splitting to fewer hierarchies, it remains the diversity of data within each hierarchy, thereby, prevent overfitting to specific data.

4.4 Discussion

Communication, privacy and computation As pointed out in [53], "privacy and communication efficiency are two primary concerns in FL". Our proposed FedDQC framework does not compromise on either of these aspects, as it only introduces an extra scalar threshold alongside the initial global model in the first round as an additional scalar parameter beyond the model parameters used in FedAvg, which is minimal. Regarding computation, FedDQC adds only one step compared to FedAvg: scoring all the training data, which only requires inferencing rather than training. When keeping the batch size the same for training and inference, the scoring time accounts for approximately 1% of the total training time. See Section 5.3.2.

Comparisons with current methods Compared to NUGGETS [50] and AlpaGasus [21], which utilize an external model for quality evaluation, FedDQC evaluates the data on the client side and preserves local data privacy. Unlike DataInf [22] and NUGGETS [50], which require an extra validation set from the server, these methods become inapplicable in scenarios where the server cannot provide this set. Additionally, their computational cost is related to the size of the validation set. Compared to IFD [25], FedDQC does not require extra dataset adaptation training, thus, is computation effective.

Algorithm 1 FedDQC: Federated Data Quality Control

Initialization: Initial global model: θ^0; Training datasets: $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$; Number of training rounds: R; Number of hierarchies: K; Global quality threshold: λ
// Scoring & Selecting Stage:
Distribute initial global model θ^0 to each client n
for $n = 1$ to N **do**
  $\mathcal{S}_n = \{s_i : s_i = f_{IRA}((q^i, a^i) \in \mathcal{D}_n, \theta^0)\}$ ▷ Assess data quality of $\mathcal{D}_n$
  $\mathcal{D}'_n = \{(q^i, a^i) \in \mathcal{D}_n, s_i \geq \lambda\}$ ▷ Select data points with quality scores above λ
  Sort $\mathcal{D}'_n$ by quality scores s_i in descending order
  Split sorted $\mathcal{D}'_n$ into hierarchies $\mathcal{H}_{n1}, \mathcal{H}_{n2}, \ldots, \mathcal{H}_{nK}$ with equal size $floor(|\mathcal{D}'_n|/K)$
  ▷ Split local dataset to hierarchies
end for
// Training Stage:
for $k = 1$ to K **do**
  for $r = (R/K) * (k-1) + 1$ to $(R/K) * k$ **do**
    for $n = 1$ to N **do**
      Local update θ_n^r with $\mathcal{H}_{nk}$ ▷ Local easy-to-hard hierarchical training
    end for
  end for
  $\theta^{r+1} = \sum_{n=1}^{N} w_n \theta_n^{r,t}$ ▷ Aggregate local models to update global model θ^r
end for
Return: Global model θ^R

5 Experiments

5.1 Experiment Setup

Dataset and evaluation metric We explore a general dataset Alpaca-GPT4 [28] and four task-specific datasets, PubMedQA [54], FiQA [55], AQUA-RAT [56] and Mol-Instructions [57] covering diverse domains (i.e., medical, finance, math, and molecular science). We apply the accuracy as the evaluation metric for PubMedQA and AQUA-RAT datasets, the BertScore [58] for Mol-Instructions dataset, the GPT-4 comparison win-rate for FiQA and the MT-Bench score for Alpaca-GPT4. For more details please refer to Appendix A.1. To demonstrate and imitate the mixed-quality data in the real world, we constructed synthetic low-quality data on four domain-specific datasets with a proportion of 50%. The low-quality data we construct needs to be challenging for data cleansing and have a significant impact on performance. Therefore, we adopted a method of constructing low-quality data by swapping answers, simulating the scenario of incorrect data responses in real situations. Additionally, this construction method also maintains the content invariance of the corpus. Examples are presented in Appendix A.5.

Models and training settings Our experiment is implemented on the OpenFedLLM [13] framework. We use LLama2-7b[3] as the pre-trained model and adapt Low-Rank Adaptation (LoRA) [59] to achieve fine-tuning. All the experiments are conducted on machines with the same hardware configuration using one NVIDIA GeForce RTX 4090. In all experiments, we use 8-bit quantization with batch size equal to 16, max length equal to 1024, and LoRA rank equal to 64 with a constant $\alpha = 128$. For the federated setting, we consider 100 communication rounds, 5 clients with $8k$ training data in total for domain-specific dataset and 20 clients with $20k$ training data in total for Alpaca-GPT4 dataset. We randomly sample 2 clients for each round with 10 local steps using AdamW [60] optimizer of model training. This setting is equivalent to 3 epochs for local training. For the NIID setting, we follow the Dirichlet distribution (with hyperparameter set to 5 for PubmedQA and FiQA, and 3 for AQUA-RAT and Mol-Instructions). We apply a cosine learning rate schedule according to the round index. The initial learning rate in the first round is $1e-4$, and the final learning rate in the last round is $1e-6$. We use the Alpaca template [61] to format the instruction, as shown in Appendix A.2.

Baselines We include four types of data quality evaluation metrics as data quality control baselines: perplexity (PPL) [62], loss, IFD [25], NUGGETS [50], and DataInf [22]. These four metrics are applied at the data-scoring stage. We select the high-score data for later federated training. Perplexity evaluates how accurately a probability model can predict a sample, usually employed in pre-training

Table 1: Performance comparisons across four datasets on both IID and NIID scenarios. FedDQC achieves the best among all datasets in both IID and NIID settings and even surpasses the full clean data training in all datasets. We bold the best performance among all data quality control methods.

Training Order	**Data Quality high/low**	**Quality Evaluation Metric**	**IID**				**NIID**			
			PubMedQA Acc	**FiQA** Win Rate	**AQUA-RAT** Acc	**Mol-Instructions** BertScore	**PubMedQA** Acc	**FiQA** Win Rate	**AQUA-RAT** Acc	**Mol-Instructions** BertScore
random	high	-	0.750	-	0.299	0.812	0.747	-	0.252	0.812
	low	-	0.681	0.266	0.205	0.809	0.664	0.354	0.205	0.809
random	low	PPL	0.703	0.437	0.224	0.809	0.684	0.544	0.217	0.804
	low	DataInf	0.728	0.457	0.224	0.811	0.675	0.464	0.232	0.807
	low	IFD	0.714	0.622	0.244	0.812	0.699	0.664	0.275	0.815
	low	NUGGETS	0.708	0.565	0.240	0.815	0.682	0.566	0.232	0.814
hierarchical	low	IRA	**0.751**	**0.709**	**0.287**	**0.822**	**0.758**	**0.794**	**0.280**	**0.823**

Table 2: Performance comparisons between various data selection baselines on the Alpaca-GPT4 [28] dataset (IID setting). FedDQC performs the best on open-ended-question benchmark MT-Bench [63].

	FedAvg	**Fedavg+**					**FedDQC**
Quality Evalaution Metric	-	Random	PPL	DataInf	IFD	NUGGETS	IRA
MT-Bench	4.78	4.56	4.58	4.76	4.65	4.38	**4.96**

data cleansing. IFD is a loss-based heuristic quality evaluation metric, that requires additional training to the specific dataset before scoring. NUGGETS define in-context prompting ability as data quality, and DataInf is an influence function adaptation to the generation tasks. In our experiments, DataInf and IFD are slightly adapted to federated scenarios, refer to Appendix A.3 for more details.

5.2 Main result

Applicability on domain-specific dataset We conduct experiments on four domain-specific datasets with synthetic low-quality data in both IID and NIID settings. We compare the FedAvg with the original dataset (referred to high-quality dataset), the synthetic mixed-quality dataset (referred to low-quality dataset), applying 4 data selection baselines and the FedDQC. For FiQA datasets, all results are compared with high-quality data with FedAvg in both settings. To fairly compare all quality evaluation metrics, we adjust the global threshold λ to guarantee the number of training data is the same in all baselines. From Table 1, we could see 1) FedDQC effectively minimizes the misaligned data influence, from the results that FedDQC consistently performs best in all datasets and both settings. 2) When comparing IID and NIID settings, data heterogeneity would slightly affect the global model performance. But this performance drop varies from dataset, for example in FiQA and PubMedQA datasets, in the NIID setting, the heterogeneity affects the global performance before data quality control. 3) FedDQC outperforms the full clean data performance in some settings. There are two reasons. The first is that progressive training methods enable the model to learn well. Secondly, even though we call the data clean, it may still consider some relatively low-quality data, negatively affecting the model performance.

Applicability on general dataset We experimented on the Alpaca-GPT4 [28] dataset in IID setting to show the effectiveness of FedDQC on general dataset. Table 2 shows the performance of full data training and performance after applying four data selection baselines and FedDQC. For all data quality control methods, we select 85% of the original data in each client based on the quality evaluation metric. We evaluate the training performance on an open-ended-question benchmark MT-Bench [63]. From Table 2, we see that (1) FedDQC significantly outperforms other data quality control methods, demonstrating its effectiveness in enhancing the general instruction tuning task's performance. (2) NUGGETS performed the worst among the baselines, even underperforming random selection. This issue may stem from the limited validation set from the server, we designed, which likely resulted in data selections that do not perform well across a sufficiently diverse dataset.

5.3 Emperical analysis of FedDQC

5.3.1 The effectiveness of hierarchical training

To demonstrate the hierarchical training mechanism's effectiveness, we compare random training, training from high-score data to low-score data, and training from low-score data to high-score data on four domain-specific datasets in IID setting, referred in the Table 3 as random, forward and reverse respectively. The experiments show that 1) The relationship between the IRA and easy-to-hard

Table 3: Performance comparisons of random batching and two hierarchical training sequences with different quality evaluation metrics on PubMedQA in IID setting. IRA is a training-aware quality evaluation metric compatible with high-to-low hierarchical training. The red box highlights the best result among all baselines, while the blue box highlights the best performance within the baseline.

	PubMedQA			AQUA-RAT			Mol-Instructions			FiQA		
Train order	**random**	**forward**	**reverse**	**random**	**forward**	**reverse**	**random**	**forward**	**reverse**	**random**	**forward**	**reverse**
random		0.681			0.205			0.809			26.60	
PPL	**0.703**	0.663	0.685	**0.240**	0.217	0.220	**0.809**	0.809	0.807	**0.437**	0.338	0.333
NUGGETS	**0.708**	0.682	0.674	**0.240**	0.193	0.201	**0.815**	0.814	0.810	0.457	**0.681**	0.320
IFD	**0.714**	0.697	0.656	**0.244**	0.217	0.193	0.814	**0.820**	0.799	**0.622**	0.612	0.287
DataInf	**0.728**	0.720	0.717	**0.224**	0.181	0.169	**0.811**	0.806	0.810	**0.565**	0.223	0.300
IRA	0.725	0.718	**0.751**	0.252	0.197	**0.287**	0.817	0.803	**0.822**	0.690	0.432	**0.709**

hierarchical training is closely intertwined, with each aspect mutually reinforcing the other. This synergy is evident in experiments where, compared to random training sequences, the application of reverse sequence training hierarchies led to notable improvements in IRA selection methods across all datasets. Notably, IRA quality selection consistently outperformed other baselines, irrespective of the training sequence. 2) The other quality evaluation metrics do not consistently benefit from the hierarchical training in all datasets, indicating its incompatibility with this hierarchical training.

5.3.2 Computational analysis

We evaluated the additional computational costs of four data quality evaluation metrics compared to IRA during the data scoring stage, alongside their training performance on the PubMedQA dataset under an IID setting in Figure 2. The experiment shows that (1) compared to the total training time in FedAvg, 300.6 minutes, IRA only takes 1% training time for data scoring, making it scalable for large datasets. (2) Compared to PPL, which is too simple to be effective. IRA uses an extra 1 minute, around 0.3% training time, for scoring than PPL but has much higher performance. (3) Compared to the second well-performed metric, DataInf, IRA takes extremely less time, around 1/150 of the scoring time than DataInf. In conclusion, IRA is a computationally efficient, scalable data quality measuring metric that could greatly enhance data quality control.

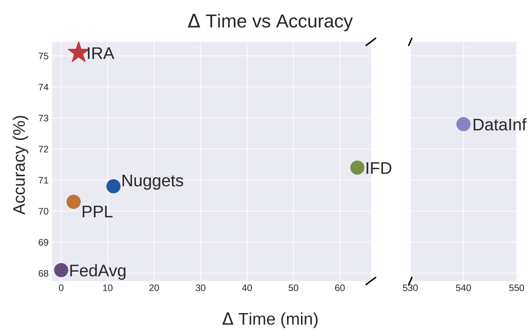


Figure 2: Comparison of additional computation costs and performance gain after applying to different quality evaluation metrics on PubMedQA [54] dataset IID setting. IRA adds minimal computational overhead while significantly enhancing performance through effective data quality control.

5.3.3 Data quality impact analysis

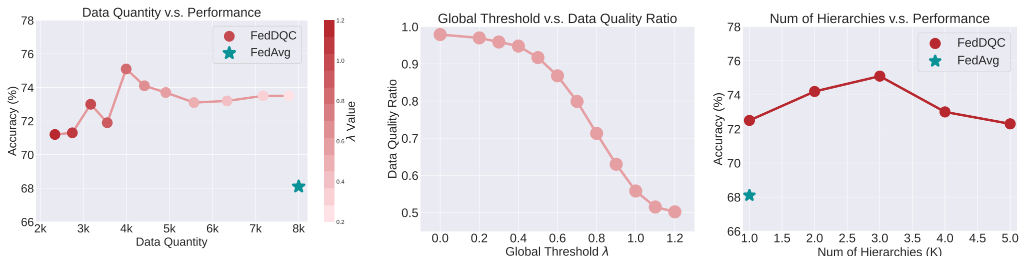
To study how data quality affects training performance, we quantify the dataset's overall quality as the ratio between the number of aligned data and the total data. We conduct experiments with different data quality ratios from 0.5 to full original data on four domain-specific datasets in the IID setting. For FiQA, we use the win rate compared to the original data set trained with FedAvg, therefore, we do not include the point when the quality ratio equals to 1.0 for FedAvg. From Fig 3, we observed (1) FedAvg performance consistently drops as the quality ratio decreases in all datasets. This phenomenon shows that low-quality data significantly affects performance, and the instruction-tuning performance proportionally relates to the data quality ratio. (2) FedDQC consistently outperforms FedAvg under all data quality ratio settings, demonstrating the effectiveness and robustness of FedDQC's data quality control. (3) Even when the data quality ratio is 1.0, meaning the training dataset does not contain synthetic low-quality data, FedDQC outperforms FedAvg across all datasets. This indicates that FedDQC can enhance training even in non-synthetic datasets, indicating its effectiveness.

5.3.4 Hyperparameter ablation

Global threshold To demonstrate the threshold robustness of FedDQCC, we further examine the impact of the global threshold λ on the PubMedQA dataset with the IID setting. Figure 4(a) shows the relationship between total data quantity used for training and performance as the global

(a) PubMedQA (b) FiQA (c) AQUA_RAT (d) Mol-Instructions

Figure 3: Comparison of FedAvg and FedDQC in various data quality ratios. (a)-(d) show the performance under different data quality ratio on PubMedQA [54], FiQA [55], AQUA-RAT [56], and Mol-Instructions [57] datasets respectively. FedDQC is consistently higher than FedAvg in all data quality ratio on four datasets.



(a) Global threshold λ (b) Global threshold v.s. Data quality ratio (c) Number of hierarchies K

Figure 4: Ablation study. (a) Effect of global threshold on overall data quantity and training performance of FedDQC. Experiments show that FedDQC is robust to the global threshold. (b) Effects of global threshold on the quality ratio of all training data. (c) The effect of various hierarchies on training performance in FedDQC training.

threshold λ is adjusted. Different shades of color of the point indicate varying λ values. The graph demonstrates that (1) as the threshold λ changes, the performance of FedDQC remains relatively stable suggesting that FedDQC is insensitive to the threshold λ. (2) Even with varying data quantities, FedDQC consistently outperforms FedAvg, indicating the robustness and effectiveness of the FedDQC approach in maintaining high performance regardless of the threshold used. (3) As Figure 4(b) shows, the data quality ratio in the selected data approaches 1.0 with a smaller threshold. Consequently, as data quantity decreases on the left side, the data quality ratio increases, but the drop in performance is more severe compared to the right side. This asymmetric performance decay around the point with 4k training data indicates that performance is more sensitive to data quantity than data quality.

Number of hierarchies Under the IID setting on PubMedQA, we tune the number of hierarchies in FedDQC $K \in \{1, 2, 3, 4, 5\}$. From Figure 4(c), we see that (1) generally $K = 3$ can lead to better performance. (2) Beyond $K = 3$ further increasing the number of hierarchies leads to a slight decline in accuracy. This suggests that while hierarchical training enhances learning by structuring data from simple to complex, too many hierarchies may reduce diversity, slightly hindering performance.

6 Conclusions and Future works

In this paper, we pioneer the exploration of data quality control in federated instruction-tuning of LLMs. We introduce a novel FL framework, Federated Data Quality Control (FedDQC), which incorporates a new data quality evaluation method, IRA, and integrates this metric with an easy-to-hard hierarchical training. FedDQC comprises two principal components: alignment-based quality assessment and quality-aware hierarchical training. Our extensive experiments demonstrate that FedDQC adds minimal computational overhead while significantly boosting performance through effective data quality control. The integration of IRA and hierarchical training is delicate and exhibits threshold robustness. However, this study did not incorporate a design aimed at enhancing data diversity. We believe that FedDQC can inspire future work focusing on integrating data quality with training processes and include diversity aspects in designation.

References

[1] Konstantinos I Roumeliotis and Nikolaos D Tselikas. Chatgpt and open-ai models: A preliminary review. *Future Internet*, 15(6):192, 2023.

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

[4] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[5] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[6] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[7] 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. *arXiv preprint arXiv:2001.08361*, 2020.

[8] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[9] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[10] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[11] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[12] Chaochao Chen, Xiaohua Feng, Jun Zhou, Jianwei Yin, and Xiaolin Zheng. Federated large language model: A position paper. *arXiv preprint arXiv:2307.08925*, 2023.

[13] Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. *arXiv preprint arXiv:2402.06954*, 2024.

[14] Momina Shaheen, Muhammad Shoaib Farooq, Tariq Umer, and Byung-Seo Kim. Applications of federated learning; taxonomy, challenges, and research trends. *Electronics*, 11(4):670, 2022.

[15] Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*, 2023.

[16] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*, 2023.

[17] Yihan Cao, Yanbin Kang, and Lichao Sun. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*, 2023.

[18] Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. *arXiv preprint arXiv:2305.09246*, 2023.

[19] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*, 2024.

[21] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

[22] Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*, 2023.

[23] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

[24] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[25] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*, 2023.

[26] Qianlong Du, Chengqing Zong, and Jiajun Zhang. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*, 2023.

[27] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[28] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[29] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[30] Anran Li, Lan Zhang, Juntao Tan, Yaxuan Qin, Junhao Wang, and Xiang-Yang Li. Sample-level data selection for federated learning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[31] Yifeng Jiang, Weiwen Zhang, and Yanxi Chen. Data quality detection mechanism against label flipping attacks in federated learning. *IEEE Transactions on Information Forensics and Security*, 18:1625–1637, 2023.

[32] Miao Yang, Hua Qian, Ximin Wang, Yong Zhou, and Hongbin Zhu. Client selection for federated learning with label noise. *IEEE Transactions on Vehicular Technology*, 71(2):2193–2197, 2021.

[33] Seunghan Yang, Hyoungseob Park, Junyoung Byun, and Changick Kim. Robust federated learning with noisy labels. *IEEE Intelligent Systems*, 37(2):35–43, 2022.

[34] Zhuowei Wang, Tianyi Zhou, Guodong Long, Bo Han, and Jing Jiang. Fednoil: A simple two-level sampling method for federated learning with noisy labels. *arXiv preprint arXiv:2205.10110*, 2022.

[35] Jingyi Xu, Zihan Chen, Tony QS Quek, and Kai Fong Ernest Chong. Fedcorr: Multi-stage federated learning for label noise correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10184–10193, 2022.

[36] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10081, 2022.

[37] Lei Wang, Jieming Bian, and Jie Xu. Federated learning with instance-dependent noisy label. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8916–8920. IEEE, 2024.

[38] Chenrui Wu, Zexi Li, Fangxin Wang, and Chao Wu. Learning cautiously in federated learning with noisy and heterogeneous clients. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 660–665. IEEE, 2023.

[39] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

[40] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, pages 1–53, 2024.

[41] Robert F Ling. Residuals and influence in regression, 1984.

[42] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[43] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *International Conference on Machine Learning*, pages 9525–9587. PMLR, 2022.

[44] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31, 2018.

[45] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.

[46] Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.

[47] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

[48] Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[49] Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. Selectit: Selective instruction tuning for large language models via uncertainty-aware self-reflection. *arXiv preprint arXiv:2402.16705*, 2024.

[50] Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, et al. One shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*, 2023.

[51] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.

[52] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[53] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[54] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[55] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.

[56] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.

[57] Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023.

[58] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[59] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[61] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[62] Javier De la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Marıa Grandury. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *arXiv preprint arXiv:2207.06814*, 2022.

[63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[64] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

[65] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*, 2023.

[66] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[67] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[68] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019.

[69] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021.

[70] Jörg Frohberg and Frank Binder. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, 2022.

[71] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[72] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

A Appendix

A.1 Dataset and Evaluation Metric

Table 4 shows descriptions of these datasets, including information about the domain, evaluation metrics, number of samples, average length of instruction, and average length of response.

Table 4: Dataset information and evaluation metrics

Dataset	Evaluation metrics	Domain	$\#samples$	$\hat{L}_{inst.}$	$\hat{L}_{Resp.}$
PubMedQA [54]	Acc	medical	211 k	471.1	71.4
FiQA [55]	Win rate	financial	17.1 k	42.1	255.7
AQUA-RAT [56]	Acc	math	97.5 k	77.4	105.7
Mol-Instructions [57]	Bert score	molecular	38 k	110.5	107.8
Alpaca-GPT4 [28]	-	general	52 k	21	163

PubMedQA PubMedQA[1] [54] is a multiple-choice question-answering dataset optimized for medical reasoning. In this paper we utilize the version sourced from PMC-LLama [64]. It features enhanced QA pairs with structured explanations derived from ChatGPT [1], facilitating in-depth medical analysis. PubMedQA dataset consists of 211.3k training samples.

FiQA FiQA dataset[2] is a subset from FinGPT [55], which consists 17.1k financial open question-answers. We split out 200 samples for evaluation and adopted the MT-Bench instruction template (see Table 5) to call ChatGPT [1] API. For the evaluation metric, we utilize the win rate to demonstrate the data quality ratio: $win\_rate = win\_counts/(win\_counts + lose\_counts)$.

Table 5: Alpaca Template for federated instruction tuning

[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. Don't provide your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.
[User Question]
{question}
[The Start of Assistant A's Answer]
{answer_a}
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]
{answer_b}
[The End of Assistant B's Answer]

AQUA_RAT The AQUA-RAT [56] dataset[3] is a large-scale mathematical dataset with a collection of around 100k algebraic word problems. Each problem in the dataset is accompanied by a detailed, step-by-step solution narrative, articulated in natural language. This dataset consists of 97.5k training samples and 245 test samples. We use accuracy as the evaluation metric.

[1] https://huggingface.co/datasets/axiong/pmc_llama_instructions
[2] https://huggingface.co/datasets/FinGPT/fingpt-fiqa_qa
[3] https://huggingface.co/datasets/aqua_rat

Mol-Instructions The Mol-Instructions [57] dataset [4] consists of a subset: biomolecular text instructions, specifically designed for natural language processing tasks in bioinformatics and chemoinformatics. It encompasses six distinct information extraction and question-answering (QA) tasks, structured through 53k detailed instructions. This design supports advanced NLP applications that require precise and context-specific understanding in the scientific domains of biology and chemistry. Our experiment only samples the open-qa task with 37k training set and 1k test set. For evaluation, the BertSocre [58], an automatic evaluation metric for text generation, is applied on a predefined test set of size 200.

Alpaca-GPT4 The Alpaca-GPT4 dataset [61] utilizes a self-instruct method to extract instructional data from ChatGPT [1], making it a widely used resource for instruction tuning. For our evaluations, we distinguish between two benchmark categories: close-ended and open-ended. The close-ended benchmarks [65] we employed include MMLU [66] for knowledge, BBH [67] and DROP [68] for reasoning, HumanEval [69] for coding, and CRASS [70] for counterfactual scenarios. For open-ended evaluation, we use Vicuna-Bench [71] and MT-Bench [63], with the latter being particularly notable for its common application in assessing instruction-following capabilities through two-turn conversation tasks.

A.2 Prompt Template

Table 6: Alpaca Template for federated instruction tuning

```
Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
{Instruction}

### Response:
```

A.3 Baselines

Perplexity: Perplexity, a probability-based metric, is defined as the exponentiated average of the negative log-likelihoods of a tokenized sequence $X = (x_0, x_1, \ldots, x_t)$. Specifically, the perplexity of X, denoted as $\mathrm{PPL}(X)$, is calculated using the formula $\mathrm{PPL}(X) = \exp\left\{-\sum_i^t \log p_\theta(x_i \mid x_{<i})/t\right\}$, where $\log p_\theta(x_i \mid x_{<i})$ represents the log-likelihood of the ith token, conditional on its preceding tokens $x_{<i}$. This measure is frequently employed to data cleaning within a pre-trained corpus [72].

DataInf Influence functions, a gradient-based scoring method, rely on the model's performance on a validation set. DataInf, as introduced by [22], stands out as the first computationally efficient approximation of influence functions that can be practically implemented in LLMs. This Hessian-based standard influence functions, provide scores $\mathrm{DataInf}(x_j)_i = \nabla L(x_j;\theta^\star) H_{\theta^\star}^{-1} \nabla L(x_i;\theta^\star)$ for every x_i in $\mathcal{D}_k$ and x_j in $\mathcal{D}_{val}$, where $\theta^\star$ denotes the parameters of the model trained on the training dataset, and $H_{\theta^\star}$ is the Hessian matrix of the empirical loss function. However, this method needs the model's convergence, which is unreal. To adapt to a federated setting, we first use the full dataset trained for 100 rounds for domain-specific datasets and 200 rounds for the general dataset. Then using this well-trained model to estimate the data influence score.

IFD The Instruction-Following Difficulty (IFD) metric is calculated by the formula $\mathrm{IFD}_\theta(Q, A) = \frac{s_\theta(A|Q)}{s_\theta(A)}$, where $s_\theta(A) = -\frac{1}{N}\sum_{i=1}^N logP(w_i^A|w_1^A, ..., w_{i-1}^A;\theta), s_\theta(A|Q) =$

[4] https://huggingface.co/datasets/zjunlp/Mol-Instructions

Table 7: Comparison between the performance of high-quality data and low-quality data according to the IRA metric.

	PubMedQA Acc	AQUA-RAT Acc	Mol-Instructions Acc	FiQA Win rate
Full data	0.750	0.2992	0.812	-
High-score	0.73	0.2559	0.822	0.7810
Low-score	0.723	0.1732	0.800	0.3733

$-\frac{1}{N}\sum_{i=1}^{N} logP(w_i^A|Q, w_1^A, ..., w_{i-1}^A; \theta)$. IFD metric measures the difficulty of following instructions of a given sample. We train our model for 20 rounds on the targeted dataset, and subsequently, this pre-trained model is used for experiments with IFD as the scoring metric.

NUGGETS NUGGETS leverages the disparity between one-shot and zero-shot scores to calculate a definitive gold score for each instruction. Exploiting the inherent contextual learning capabilities of large models.

A.4 How Data quality affects training performance

We compare the high-score proportion of data with the low-score proportion of data and show that the data quality indeed affects training performance. See Table 7.

A.5 Case Study

A.5.1 Examples of synthetic low-quality data

Table 8, 9, 10, 11 shows examples of synthetic low-quality data in 4 domain-specific datasets. Each dataset showcases a pair of data, including the questions, the authentic correct answers, and the incorrect answers generated by swapping the correct ones. The synthetic low-quality data is created by pairing the same question with its incorrect answer.

A.5.2 Examples of scored data

Figures 5, 6, 7, 8 shows examples of scored data in 4 mixed-quality domain-specific datasets. Each dataset's size is 8k, with 50% low-quality data generated by swapping correct answers. The remaining 50% is considered high-quality data. We use IRA as the scoring metric and show typical data examples with scores in top 1% and lowest 1%.

Typically, high-quality data scores high and low-quality data scores low. This is because the incorrect answers in low-quality data significantly diminish the instruction-response relativeness, leading to an increase in IRA. However, the high-quality data example in Figure 5 scores low, due to the presence of complicated and verbose input. Consequently, the model finds it challenging to establish the relativeness between the instruction and response.

Table 8: Synthetic low-quality data in PubMedQA

Question	**Correct Answer**	**Incorrect Answer**
Given your profession as a doctor, please provide responses to the medical questions using the patient's description. Analyze the question given its context. Give both long answer and yes/no decision. ###Question: Does mineralocorticoid receptor blockade inhibit accelerated atherosclerosis induced by a low sodium diet in apolipoprotein E-deficient mice? ###Context: A low-sodium diet (LSD) was shown to increase both angiotensin II (AngII) and aldosterone levels ...	In E0 mice, Epl inhibited LSD-accelerated atherosclerosis, despite the elevation of renin and aldosterone levels. It is therefore suggested that the atherogenic action of LSD could be mediated, at least in part, by activation of the mineralocorticoid receptor. In addition, eplerenone may have direct anti-inflammatory actions. ###Answer: yes	Therefore, the technique could be considered an alternative method in increasing the keratinized tissue for cases that have a minimal amount of keratinized tissue. ###Answer: yes
In your capacity as a doctor, it is expected that you answer the medical questions relying on the patient's description. Analyze the question given its context. Give both long answer and yes/no decision. ###Question: Do simultaneous placement of an interpositional free gingival graft with nonsubmerged implant placement? ###Context: Various surgical techniques target achieving adequate keratinized tissue around dental implants ...	Therefore, the technique could be considered an alternative method in increasing the keratinized tissue for cases that have a minimal amount of keratinized tissue. ###Answer: yes	In E0 mice, Epl inhibited LSD-accelerated atherosclerosis, despite the elevation of renin and aldosterone levels. It is therefore suggested that the atherogenic action of LSD could be mediated, at least in part, by activation of the mineralocorticoid receptor. In addition, eplerenone may have direct anti-inflammatory actions. ###Answer: yes

Table 9: Synthetic low-quality data in FiQA

Question	Correct Answer	Incorrect Answer
Share your insights or perspective on the financial matter presented in the input. ###Question: What one bit of financial advice do you wish you could've given yourself five years ago?",	Compound interest. Next time you buy a 100$ toy realize that if you save it - in x years that 100$ you saved and invested could potentially be more than 100$ where as most likely whatever you're buying will be worth much less.	If the building has no income, it also probably has minimal expenses. The heat, water and electricity costs are nearly zero. They are letting the value depreciate, and taking it off the taxes. I also suspect the condition of the building is poor, so any effort to make the building productive would be very costly. Many cities combat this by setting the tax on empty buildings or empty lots at a much higher rate. ...
Utilize your financial knowledge, give your answer or opinion to the input question or subject . Answer format is not limited. ###Question: Tax deductions on empty property	If the building has no income, it also probably has minimal expenses. The heat, water and electricity costs are nearly zero. They are letting the value depreciate, and taking it off the taxes. I also suspect the condition of the building is poor, so any effort to make the building productive would be very costly. Many cities combat this by setting the tax on empty buildings or empty lots at a much higher rate. ...	Compound interest. Next time you buy a 100$ toy realize that if you save it - in x years that 100$ you saved and invested could potentially be more than 100$ where as most likely whatever you're buying will be worth much less.

Table 10: Synthetic low-quality data in AQUA-RAT

Question	Correct Answer	Incorrect Answer
###Instruction: Choose the correct option for the following math question. ###Question: 1000 men have provisions for 15 days. If 200 more men join them, for how many days will the provisions last now? ###Options: A. 12.8 B. 12.4 C. 12.5 D. 16.8 E. 92.7	###Rationale: 1000*15 = 1200*x x = 12.5 ###Answer: OPTION C IS CORRECT.	###Rationale: Explanation: Let the sum of money be x then (x × 4 × 8)/100 = (560 × 12 × 8)/100 x × 4 × 8 = 560 × 12 × 8 x × 4 = 560 × 12 x = 560 × 3 = 1680 ###Answer: OPTION D IS CORRECT.
###Instruction: Choose the correct option for the following math question. ###Question: If simple interest on a certain sum of money for 8 years at 4% per annum is same as the simple interest on Rs. 560 for 8 years at the rate of 12% per annum then the sum of money is ###Options: A. Rs.1820 B. Rs.1040 C. Rs.1120 D. Rs.1680 E. None of these	###Rationale: Explanation: Let the sum of money be x then (x × 4 × 8)/100 = (560 × 12 × 8)/100 x × 4 × 8 = 560 × 12 × 8 x × 4 = 560 × 12 x = 560 × 3 = 1680 ###Answer: OPTION D IS CORRECT.	###Rationale: 1000*15 = 1200*x x = 12.5 ###Answer: OPTION C IS CORRECT.

Table 11: Synthetic low-quality data in Mol-Instructions

Question	Correct Answer	Incorrect Answer
Answer this question truthfully ###Question: What is the predicted relative molecular mass of the protein encoded by PVAS2?	The predicted relative molecular mass of the protein encoded by PVAS2 is 65810 Da.	Resonance Raman spectroscopy is a form of spectroscopy used to analyze the vibrational, rotational, and other structural characteristics of molecules.
Answer this question truthfully ###Question: What is Resonance Raman spectroscopy?	Resonance Raman spectroscopy is a form of spectroscopy used to analyze the vibrational, rotational, and other structural characteristics of molecules.	The predicted relative molecular mass of the protein encoded by PVAS2 is 65810 Da.

Scored data examples in PubMedQA

High-quality, High-score
IRA score: 4.08

[Instruction]
Considering your role as a medical practitioner, please use the patient's description to answer the medical questions.
Analyze the question given its context. Give both long answer and yes/no decision.
[Input]
###Question: Does [Hemorrhagic shock increase the occurrence of bacterial translocation]?
###Context: To determine whether hemorrhagic shock (HS) increases the occurrence of bacterial translocation (BT). 100 patients were divided into 4 groups: control group (group I, 34 patients); group with hemorrhagic shock (HS) caused by closed blunt abdominal trauma (group II, 23); group caused by closed blunt abdominal trauma without HS (group III 15); and group with HS caused by intra-abdominal viscus hemorrhage (group IV 28). Preoperative and postoperative samples were taken from peripheral blood, visceral peritoneal swab, portal vein blood, ileal mesenteric lymph node, liver and spleen biopsy respectively for aerobic and anaerobic culture. The positive culture rates of these groups were 6%, 65%, 13%, 68% respectively. The difference between the control and experimental groups was significant($P < 0.05$). The difference was also significant between group I and II and between I and IV ($P < 0.01$), whereas it was not significant between I and III, and between II and IV ($P > 0.05$).
[Output]
HS increases the occurrence of BT.

###Answer: yes

High-quality, Low-score
IRA score: -0.61

[Instruction]
Given your background as a doctor, please provide your insight in addressing the medical questions based on the patient's account.
Analyze the question given its context. Give both long answer and yes/no decision.
[Input]
###Question: Does globulin-platelet model predict minimal fibrosis and cirrhosis in chronic hepatitis B virus infected patients?
###Context: To establish a simple model consisting of the routine laboratory variables to predict both minimal fibrosis and cirrhosis in chronic hepatitis B virus (HBV)-infected patients. We retrospectively investigated 114 chronic HBV-infected patients who underwent liver biopsy in two different hospitals. Thirteen parameters were analyzed by step-wise regression analysis and correlation analysis. A new fibrosis index [globulin/platelet (GP) model] was developed, including globulin (GLOB) and platelet count (PLT). GP model = GLOB (g/mL) × 100/PLT (× 10(9)/L). We evaluated the receiver operating characteristics analysis used to predict minimal fibrosis and compared six other available models. Thirteen clinical biochemical and hematological variables [sex, age, PLT, alanine aminotransferase, aspartate aminotransferase (AST), albumin, GLOB, total bilirubin (T.bil), direct bilirubin (D.bil), glutamyltransferase, alkaline phosphatase, HBV DNA and prothrombin time (PT)] were analyzed according to three stages of liver fibrosis (F0-F1, F2-F3 and F4). Bivariate Spearman's rank correlation analysis showed that six variables, including age, PLT, T.bil, D.bil, GLOB and PT, were correlated with the three fibrosis stages (FS). Correlation coefficients were ...
[Output]
By establishing a simple model using available laboratory variables, chronic HBV-infected patients with minimal fibrosis and cirrhosis can be diagnosed accurately, and the clinical application of this model may reduce the need for liver biopsy in HBV-infected patients.

###Answer: yes

Figure 5: The examples of scored data in mixed-quality PubMedQA dataset include a high-quality datapoint with a high score and a high-quality datapoint with a low score.

Scored data examples in FiQA

High-quality, High-score
IRA score: 3.16

[Instruction]
Share your insights or perspective on the financial matter presented in the input.
[Input]
###Question: On a debt collection agency's letter, what does "balance" refer to?
[Output]
The balance is the amount due.

Low-quality, Low-score
IRA score: -0.28

[Instruction]
Utilize your financial knowledge, give your answer or opinion to the input question or subject . Answer format is not limited.
[Input]
###Question: How can a person with really bad credit history rent decent housing?
[Output]
I can't give you a detailed answer because I'm away from the computer where I use kMyMoney, but IIRC to add investments you have to create new transactions on the 'brokerage account' linked to your investment account.

Figure 6: The examples of scored data in mixed-quality FiQA dataset include a high-quality datapoint with a high score and a low-quality datapoint with a low score.

Scored data examples in AQUA-RAT

High-quality, High-score
IRA score: 1.92

[Instruction]
###Instruction: Choose the correct option for the following math question.
[Input]
###Question: What should come in place of question mark(?) in the following equation?
8597 – ? = 7429 – 4358
###Options:
A. 5706
B. 5526
C. 5426
D. 5626
E. None of these

[Output]
###Rationale: 8597 – ? = 7429 – 4358
? = 5526

###Answer: OPTION B IS CORRECT.

Low-quality, Low-score
IRA score: -0.04

[Instruction]
###Instruction: Choose the correct option for the following math question.
[Input]
###Question: A jar contains only red, yellow, and orange marbles. If there are 3 red, 5 yellow, and 4 orange marbles, and 2 marbles are chosen from the jar at random without replacing any of them, what is the probability that 2 yellow, 1 red, and no orange marbles will be chosen?
###Options:
A. 1/60
B. 1/45
C. 2/45
D. 3/22
E. 6/22

[Output]
###Rationale: P= 16 = 16+3 = 19 = S
O = 15 =15+3 =18 = R
SIMILARLY,
P = 16 = 16+3 = 19 = S
R = 18 = 18+3 = 21 = U
I = 9 = 9+3 = 12 = L
Y = 25 = 25 +3 = 28 = B
A = 1 = 1+3 = 4 = D
SO ANS IS
SULB

###Answer: OPTION D IS CORRECT.

Figure 7: The examples of scored data in mixed-quality AQUA-RAT dataset include a high-quality datapoint with a high score and a low-quality datapoint with a low score.

Scored data examples in Mol-Instructions	
High-quality, High-score IRA score: 3.95 [Instruction] Answer this question truthfully [Input] ###Question: How long after the beginning of the secondary aggregation process does the mitotic coefficient reach a first maximum? [Output] The mitotic coefficient reaches a first maximum 8 hours after the beginning of the secondary aggregation process.	Low-quality, Low-score IRA score: -0.63 [Instruction] Answer this question truthfully [Input] ###Question: What are system invariants? [Output] The results suggest that multiple pathways exist in soybean to regulate expression of Aox genes.

Figure 8: The examples of scored data in mixed-quality Mol-Instructions dataset include a high-quality datapoint with a high score and a low-quality datapoint with a low score.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim that we propose a new framework of federated instruction tuning of LLM with data quality control(FedDQC), which measures the data quality to facilitate the subsequent processes of filtering and hierarchical training.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include our limitation in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: NA

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our method with a lot of detail and offer enough information to reproduce the results. We also have a detailed appendix to make sure experiments are all reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We are not able to release the code until the paper is accepted. But we ensure that we will make the code and data available to the public once we get accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Section 5, we describe all the training and evaluation details, and make sure readers can follow them up.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: It's not clear what assumption should be made for the error distribution in our setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the Section 5, we clarify that one NVIDIA RTX 4090 is used for training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read the code of ethics carefully and believe we follow them strictly.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We do not discuss the social impacts since it is a federated learning framework, which in itself has at least no negative social impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not believe this framework have high risks for misuse. It should benefit the federated learning and LLM community.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all the related papers and models in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not have crowdsourcing and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not have these stuffs.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.