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

Anonymous ACL submission

Abstract

Federated Learning (FL) enables privacy-preserving collaborative instruction tuning of large language models (LLMs) by leveraging massively distributed data. However, the decentralized nature of FL exacerbates data quality challenges, as local clients lack global visibility to filter noisy or low-quality samples before training. To resolve this issue, we propose FedDQC, a novel federated instruction tuning framework with dynamic data quality control. Our approach introduces two key innovations. First, we propose instruction-response alignment (IRA)—an efficient client-side metric for quality evaluation requiring only low-cost inference. We validate that higher-IRA data corresponds to more relevant and easier-to-learn question-answer pairs. Second, mirroring the human easy-to-hard knowledge acquisition process, we design a quality-aware hierarchical FL training framework, where the LLM is progressively fine-tuned from high- to low-IRA data in a collaborative manner. The framework also supports adaptive data quality assessment at each hierarchy, enabling dynamic adjustments throughout the training process. Extensive experiments on synthetic and real-world datasets show that our method significantly improves LLM performance on mixed-quality data in FL.

1 Introduction

For large language models (LLMs) training (Roumeliotis and Tselikas, 2023; Chowdhery et al., 2023; Touvron et al., 2023; Jiang et al., 2023a), both the quantity and quality of the training data significantly impact their performance (Zhao et al., 2023; Minaee et al., 2024). The scaling law suggests that more training data can lead to more powerful LLMs (Kaplan et al., 2020). However, in specific domains such as healthcare (Thirunavukarasu et al., 2023) and finance (Wu et al., 2023c), privacy concerns (Albrecht, 2016) prevent the aggregation of large-scale datasets, making it challenging to expand the dataset scale. Federated Learning (FL) (McMahan et al., 2017), as 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 (Chen et al., 2023a; Ye et al., 2024c; Kuang et al., 2024; Fan et al., 2023; Ye et al., 2024a).

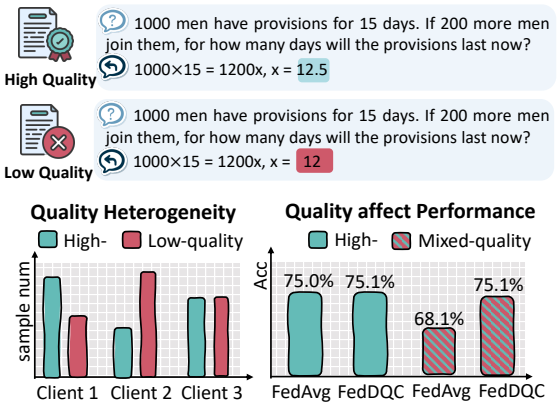


Figure 1: Top figure is an example of low-quality data and high-quality data. The left figure shows federated quality heterogeneity. The right figure shows how data quality affects federated training performance and FedDQC eliminates low-quality data effects.

While FL addresses the data quantity issue by incorporating more local clients, it brings more data quality issues (Shaheen et al., 2022). In FL, training data for each client are collected from various sources locally, making it difficult to detect low-quality data or noises in local datasets. Such vulnerabilities adversely affect general model training. Although numerous methods (Wu et al., 2023b; Liu et al., 2023; Cao et al., 2023; Chen et al., 2023b; Zhou et al., 2024) are proposed for data quality control in LLM instruction tuning, 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 focus on designing data quality evaluation metrics to quantify data quality. They can be broadly categorized into two types. The first category consists of heuristic-based methods specifically designed for instruction-tuning tasks. (Wang et al., 2024a) These methods quantify the quality of instruction-response pairs (Li et al., 2023a; Du et al., 2023; Cao et al., 2023) by quantity of information. However, they rely on the assumption that all data are clean and reliable making them difficult to deal with noises and errors in dataset. The second category is traditional data attribution methods (Kwon et al., 2023; Ghorbani and Zou, 2019; Ilyas et al., 2022), which require re-training and evaluation on the whole dataset. However, in FL each client only has limited computation resources for re-training and does not allow access to the local dataset, making these methods impractical for FL.

To fill this gap, we propose FedDQC (Federated Data Quality Control), a novel FL framework with data quality control for LLM instruction tuning. First, we propose an efficient and privacy-preserving data quality scoring metric: IRA (Instruction-Response Alignment), which could be computed on the client side with minimum cost. This metric evaluates the data quality by estimating the mutual information between the instruction and response on LLM. Specifically, it calculates the response inference loss difference between given instruction and without instruction. This approach fully leverages the knowledge embedded in the pre-trained LLM and eliminates the impact of response format inconsistency with pre-trained data. In the context of instruction tuning, higher alignment indicates that the instruction and response are better matched, enabling the model to learn how to answer questions more easily.

Based on the proposed IRA scores, we propose an FL training framework fully leveraging data quality evaluation to handle data quality issues. The key idea behind FedDQC is a combination of hierarchical training and adaptive scoring during training. Specifically, it consists of two stages: the scoring stage and the hierarchical training stage. In the scoring stage, each client independently computes IRA scores for local data samples using the current global model through localized inference without requiring external data access. All data is then re-ordered based on IRA scores, and high-quality data is selected and partitioned into hierarchical subsets for subsequent training. in the hierarchical training stage, the model initially trains on high-quality, easily learnable samples (higher IRA scores) and gradually transitions to more complex data (lower IRA scores). This staged approach reduces interference from challenging samples in early training phases, thereby enhancing learning efficiency. These two stages iterate until the final hierarchy, effectively enabling adaptively scoring during training. This progressive knowledge integration mechanism allows the model to incorporate more challenging data as its capacity improves. This leads to enhanced model robustness and performance, particularly when dealing with noisy or heterogeneous data.

Our experiments demonstrate that FedDQC not only outperforms all baseline models in both IID (independent and identically distributed) and non-IID settings on four synthetic datasets but also shows effectiveness on the real-world federated dataset, Fed-WildChat (Ye et al., 2024b). As for computation, we show that the scoring metric IRA consumes only 1% training time for data quality evaluation, making it computation-efficient and scalable for larger datasets.

2 Related work

2.1 Federated Learning

Federated Learning (Kairouz et al., 2021; McMahan et al., 2017; Li et al., 2020) 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. (Li et al., 2021) 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 (Jiang et al., 2023b; Yang et al., 2021) through feature (Yang et al., 2022) or model weight clustering (Wang et al., 2022). While at the sample level, studies have typically focused on label correction strategies (Xu et al., 2022) or confidence-based sample reweighting (Fang and Ye, 2022). At the model level, approaches like distillation (Wang et al., 2024b) or modifying the loss function (Wu et al., 2023a) 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

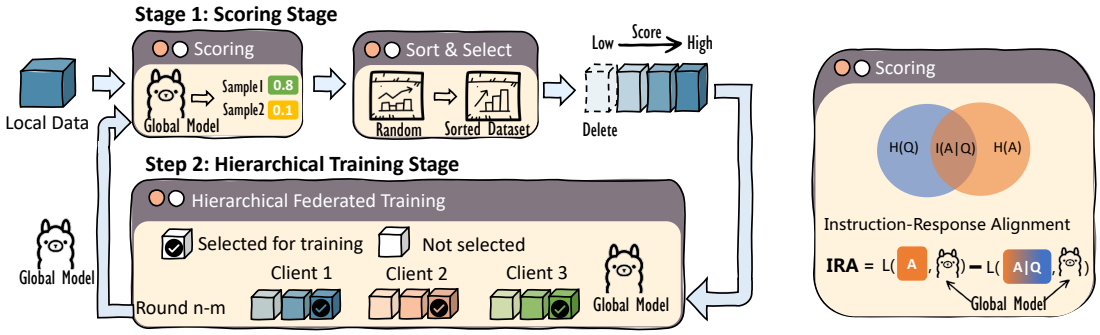


Figure 2: Overview of FedDQC, which iterates in two stages: (1) Scoring stage: utilize IRA and global model to evaluate data quality; (2) Hierarchical training: progressively fine-tuned from high-IRA to low-IRA data, mirroring the easy-to-hard learning process.

tailored to generative tasks in FL.

2.2 Data quality control

Data quality control is complex and a throughout problem in machine learning (Zha et al., 2023). 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. (Hammoudeh and Lowd, 2024) Retraining-based approaches, such as leave-one-out (Ling, 1984), Shapley value (Ghorbani and Zou, 2019), and Datamodels (Ilyas et al., 2022), estimate the effect of data points by repeatedly retraining the model on different subsets of data. These data attribution approaches are post-hoc and computationally costly, making them unsuitable for LLM setting. Gradient-based approaches, like represented point selection (Yeh et al., 2018), TracIn (Pruthi et al., 2020), and influence functions (Koh and Liang, 2017), estimate training data's impact through parameter sensitivity. Recent studies have developed more efficient adaptations of this gradient-based method for generative tasks (Guo et al., 2020) and LLM settings, streamlining data selection processes such as pre-training (Park et al., 2023) and instruction-tuning in transfer learning scenarios (Xia et al., 2024). 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 (Chen et al., 2023c), DEITA (Liu et al., 2023) and INSTAG (Lu et al., 2023) prompt ChatGPT (Roumeliotis and Tselikas, 2023) 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 (Li et al., 2023a) and MoDS (Du et al., 2023) require a computationally intensive initial training stage that may involve low-quality data. Similarly, InstructionMining (Cao et al., 2023) despite utilizing innovative statistical regression to fit quality influence factors with performance, is dataset-specific and requires retraining. Additionally, approaches like SelectIT (Liu et al., 2024) and NUGGETS (Li et al., 2023b) 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 Preliminary: Federated Learning

We consider there are N clients participating in FL to collaboratively train a model θ. Each client holds a dataset $\mathcal{D}_n$ and optimizes its local model via a loss function $L(\cdot)$. The goal of FL is to find the optimal global model θ^* that minimizes the aggregated loss of all clients. Mathematically, the global objective of FL is:

$$\theta^* = \arg\min_{\theta} \sum_{n=1}^{N} \frac{w_n}{|\mathcal{D}_n|} \sum_{x \in \mathcal{D}_n} L(x, \theta)$$

where $w_n = \frac{|\mathcal{D}_n|}{\sum_{i=1}^{N} |\mathcal{D}_i|}$ represents the weight assigned to client, $|\mathcal{D}_n|$ is dataset size of $\mathcal{D}_n$.

In the basic 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 Federated Instruction Tuning

In federated instruction tuning, each client holds a dataset where each sample is a pair: (question, answer). For client n, the dataset is denoted as $\mathcal{D}_n = \{(q^i, a^i) | i = 1, 2, \ldots, |\mathcal{D}_n|\}$, where q^i and a^i denote the i-th instruction and answer. 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.

4 Methodology

This section presents the two-stage FedDQC framework: the scoring stage and the training stage. Firstly, Section 4.1 gives an overview of FedDQC. Then Section 4.2 and Section 4.3 introduce the scoring and training stage respectively.

4.1 Overview

FedDQC operates through an iterative two-stage process: the scoring stage and the training stage. These stages alternate in a continuous cycle, allowing the model to select and progressively learn from high-quality data.

Scoring Stage: At the beginning of each hierarchy, clients assess the quality of their local data using the IRA metric, which evaluates the alignment between instructions and responses. Based on these scores, clients sort and filter their data, retaining only high-quality samples for federated training.

Training Stage: Client partition the filtered high-quality local data into several subsets based on sorted sequence, with each subset with equal size. In each hierarchy training, clients only choose the highest-scored subset to participate in federated training in this hierarchy.

Please refer to Fig. 2 and an algorithmic summary in Algorithm 1.

4.2 Scoring Stage: Data Quality Assessment

FedDQC controls data quality in FL by locally assessing data under privacy and computation constraints, allowing clients to select and sort data based on quality.

Quality evaluation metric We propose the Instruction-Response Alignment (IRA) metric for data quality assessment, inspired by mutual information (Kraskov et al., 2004), to evaluate the relevance between instructional prompts and responses using the global model. Specifically, IRA is computed locally with the global model to calculate the difference in inference loss between unconditioned responses and responses conditioned on their corresponding instructions. The following equation defines the scoring function f_{IRA}:

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

where $L(a^i; \theta) = -\sum_{j=1}^{l_i} \log p(a_j^i | a_{<j}^i; \theta)$ calculates the cross-entropy loss of generating response a^i without given the instruction q^i, $L((a^i, q^i); \theta)$ is the cross-entropy loss given instruction q^i, which is defined in Section 3.2. $\mathcal{D}$ is dataset and θ represents model parameter for data quality evaluation.

This metric subtly connects data quality with learning difficulty by reflecting how well the instruction aligns with the response, which in turn influences how easily the model can learn from the data. Visualization in Fig. 3(c) supports this, see discussion in Section 6.3.

Local dataset sort and select Using the efficient IRA, clients process the local dataset in two steps. First, clients sort their untrained local data in a descending order based on IRA values. Then client filter out low-quality samples using a global threshold λ, retaining only the high-quality data.

Rather than solely relying on the pre-trained model to evaluate data quality, our approach incorporates re-scoring before each hierarchical federated training stage, dynamically adjusting the data selection process according to the model's evolving capabilities throughout training. This adaptive approach offers a key advantage: better handling of low-quality data. As the model's abilities improve during training, it becomes more adept at distinguishing between challenging, valuable data and noisy or irrelevant data, making it more capable in selecting the right data and ensuring a more reliable and stable model. Section 6.3 provides visualizations that offer strong evidence of the effectiveness of this methodology.

4.3 Training Stage: Hierarchical Training

After controlling data quality in local dataset, the next step is federated training. In this stage, we propose quality-aware hierarchical training based on previous IRA scoring, where models learn pro-

Table 1: Performance comparisons on real and synthetic datasets in both IID and NIID settings show that FedDQC outperforms all methods and even surpasses full clean data training. The best performance for each data quality control method is bolded.

	Real Dataset	**Sythetic Dataset**							
		IID				**NIID**			
	Fed-WildChat	**PubMedQA**	**FiQA**	**AQUA-RAT**	**Mol-Instructions**	**PubMedQA**	**FiQA**	**AQUA-RAT**	**Mol-Instructions**
	MT-bench	Acc	Win%	Acc	BertScore	Acc	Win%	Acc	BertScore
FedAvg (oracle)	4.475	0.750	-	0.299	0.812	0.747	-	0.252	0.812
FedAvg	-	0.681	0.266	0.205	0.809	0.664	0.354	0.205	0.809
FedAvg+PPL	4.525	0.703	0.437	0.224	0.809	0.684	0.544	0.217	0.804
FedAvg+DataInf	4.443	0.728	0.457	0.224	0.811	0.675	0.464	0.232	0.807
FedAvg+IFD	4.600	0.714	0.622	0.244	0.812	0.699	0.664	0.275	0.815
FedAvg+NUGGETS	4.443	0.708	0.565	0.240	0.815	0.682	0.566	0.232	0.814
FedDQC	**4.780**	**0.751**	**0.721**	**0.290**	**0.819**	**0.751**	**0.821**	**0.280**	**0.824**

gressively from easier to harder data. This typically involves two steps:

Step 1: Split to subsets: For hierarchy k in total K hierarchies, the retained data for client n is partitioned into $K-k$ separate hierarchies in a descending $\mathcal{H}_{nk}, \ldots, \mathcal{H}_{nK}$ order with IRA values. Each hierarchy contains an equal number of samples, and any samples that have already been trained in previous hierarchies are removed. The subset $\mathcal{H}_{nk}$ with the highest score is selected as the training set for this hierarchical federated training.

Step 2: Federated training inside hierarchy: During training, each clients choose the highest scored subset $\mathcal{H}_{nk}$ for training. By prioritizing high-quality, easy-to-learn data, the FL process starts with basic, highly relevant instruction-response samples, then gradually applies its instruction-following skills to more generalized tasks, and eventually progresses to solving more complex problems. This approach offers two key benefits: 1) it enables the model to build a strong foundational understanding, improving learning effectiveness and robustness; 2) it ensures consistent data quality in each training round, reducing the risk of divergence in the aggregated model. See visualization in Section 6.3

In each round of global aggregation and local updating, any federated algorithm could easily adapt to FedDQC. Section 6.2 demonstrate the effectiveness of FedDQC plugged in various FL algorithms.

5 Disscussion

Communication, privacy and computation "Privacy and communication efficiency are two primary concerns in FL". (Kairouz et al., 2019) Our proposed FedDQC does not compromise on either of these aspects, as it does not introduce extra communication costs or privacy leakage, through training. 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 6.4.2.

6 Experiments

6.1 Experiment Setup

Dataset and evaluation metric We explore a real-world dataset and four task-specific datasets, PubMedQA (Jin et al., 2019), FiQA (Yang et al., 2023), AQUA-RAT (Ling et al., 2017) and Mol-Instructions (Fang et al., 2023) covering diverse domains (i.e., medical, finance, math, and molecular science). To simulate real-world mixed-quality data, we introduced synthetic low-quality data at a 50% proportion across the four domain-specific datasets. For more details please refer to Appendix A.1.

Models and training settings Our experiment is implemented on the OpenFedLLM (Ye et al., 2024c) framework. We use LLama2-7b(Touvron et al., 2023) as the pre-trained model and adapt Low-Rank Adaptation (LoRA) (Hu et al., 2021) to achieve fine-tuning. See Appendix A.4.1.

Baselines We include four types of data quality evaluation metrics as data quality control baselines: perplexity (PPL) (De la Rosa et al., 2022), loss, IFD (Li et al., 2023a), NUGGETS (Li et al., 2023b), and DataInf (Kwon et al., 2023). These four metrics are applied at the data-scoring stage. We select the high-score data for later federated training. In our experiments, DataInf and IFD are slightly adapted to federated scenarios, refer to Appendix A.3 for more details.

6.2 Main result

We conduct experiments on a real-world dataset and four synthetic domain-specific datasets with synthetic low-quality data on both IID and NIID settings, shown in Table 1.

Applicability on synthetic dataset We compare FedAvg with the original dataset (referred as oracle

Table 2: Performance comparisons of random batching and two hierarchical training sequences with 5 quality evaluation metrics on PubMedQA in IID setting. IRA is a training-aware quality evaluation metric compatible with descending 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**	**ascend**	**descend**	**random**	**ascend**	**descend**	**random**	**ascend**	**descend**	**random**	**ascend**	**descend**
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	**0.290**	0.817	0.803	**0.819**	0.690	0.432	**0.721**

in table), and the synthetic mixed-quality dataset, applying 4 data selection baselines and the FedDQC. For FiQA datasets, all win rate are compared with high-quality data with FedAvg in both settings. To ensure fairness, we adjust the global threshold λ to keep the number of training samples consistent and maintain the same number of training rounds. Key observations from Table 1 include: 1) FedDQC consistently mitigates the impact of low-quality data and outperforms other baseline. 2) In some cases, FedDQC even surpasses the performance of training on fully clean data, benefiting from progressive training and the fact that not all data in the oracle dataset is equally valuable.

Applicability on real-world dataset We present the results on the real-world dataset, Fed-WildChat, in Table 1, where various data scoring metrics were applied to select 70% of the oracle training data for the same number of training rounds. We can see that: 1) FedDQC outperforms all other baselines and even surpasses the full dataset training performance, indicating the presence of low-quality data in the real-world dataset. 2) Data selection methods based on DataInf and NUGGEST perform worse than using the full dataset. This suggests that using gradients for data attribution in real-world datasets, as in the case of DataInf, is challenging. Additionally, the diverse distribution of real-world data makes it difficult to evaluate data quality using a fixed validation set, as shown by the performance of NUGGEST.

Applicability on different FL algorithms We combine FedDQC with several FL algorithms beyond FedAvg, including FedAvgM (Hsu et al., 2019), FedAgrad (Reddi et al.), FedYOGI (Reddi et al.), and FedAdam (Reddi et al.). Table 3 shows that FedDQC significantly boosts performance across these algorithms. For example, in the mix-quality scenario, FedAdagrad's performance improved from 0.709 to 0.731 with FedDQC, illustrating the effectiveness of FedDQC in enhancing model performance when paired with other FL algorithms.

Table 3: Compatability with other 4 federated algorithms on PubMedQA, IID setting. The last line shows the improvement on mixed-quality data with FedDQC added.

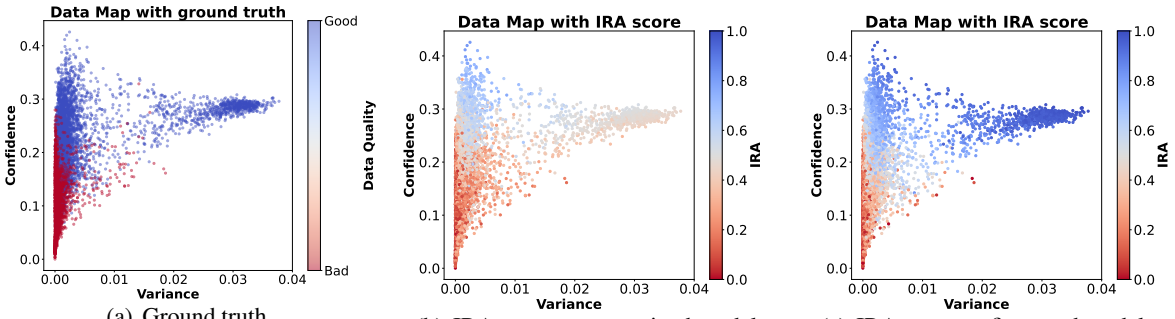
	FedAvg	FedAvgM	FedAdagrad	FedYOGI	FedAdam
oracle	0.750	0.732	0.717	0.512	0.527
mix-quality	0.681	0.676	0.709	0.498	0.476
+FedDQC	0.751	0.729	0.731	0.512	0.531
	(+7%)	**(+5.3%)**	**(+2.2%)**	**(+1.4%)**	**(+5.5%)**

6.3 Visualization

Data Map (Swayamdipta et al., 2020) is a data training dynamics visualization tool, which tracks each sample's inference probability across training epochs. The Confidence (y-axis) is the mean of these probabilities and the Variance (x-axis) is the variance of these probabilities. Fig 3 shows 5 types of low-quality data (noisy token, deleted token, truncation, and swapped responses) in a centralized setting on LLaMA-2-7b (Touvron et al., 2023) model and dataset PMC-Llama (Wu et al., 2024) with 8000 samples, of which 50% are low-quality samples.

Relation of IRA and training difficulty IRA metric subtly connects data quality with learning difficulty. Fig. 3(b) shows how this evaluation method closely reflects the relationship between IRA scores and the dynamic process of data during training. The Data Map in Fig. 3(a) reveals a clear pattern between data quality and its training dynamics. For instance, data with high confidence and high variance are easier for the model to learn and perform better on, while low-variance, low-confidence data, like those in the bottom-left corner, are harder to learn and represent low-quality data. As Fig. 3(b) illustrated, IRA aligns well with this dynamic tracking approach: 1) high-scoring data are easier to learn, with higher variance and lower confidence; 2) low-scoring data tend to cluster in the lower-left corner, with both lower confidence and variance, indicating more difficulty in learning, making them more likely to be noisy or irrelevant and negatively impacting model performance.

Effectiveness of iterative scoring As shown in Fig.3(c), the model re-scored after training dis-



(a) Ground truth (b) IRA score on pre-trained model (c) IRA score on finetuned model

Figure 3: Data Map visualization. (a) Data Map with ground truth quality label. (b) Data Map with IRA scores on a pre-trained model. (c) Data Map with IRA scores on fine-tuned model.

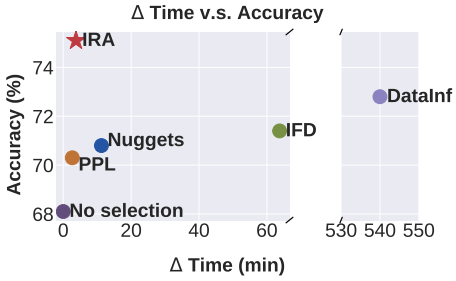
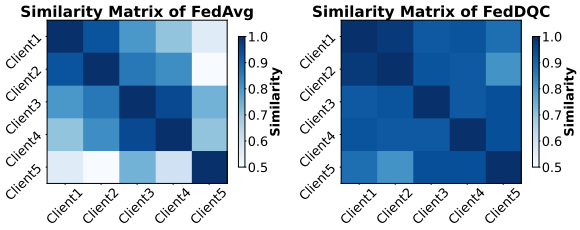


Figure 4: Comparison of additional computation costs and performance gain after applying to different quality evaluation metrics on PubMedQA IID setting. IRA adds minimal computational overhead while significantly improves performance by data quality control.



(a) FedAvg (b) FedDQC

Figure 5: Model similarity comparison between FedAvg and FedDQC on PubMedQA dataset in NIID settings.

tinguishes data quality more clearly than the pre-trained model in Fig.3(b). Notably, high-scoring data tend to appear in regions with higher variance, indicating that the model is more confident in these challenging, yet informative samples.

6.4 Emperical analysis of FedDQC

6.4.1 The effectiveness of hierarchical training

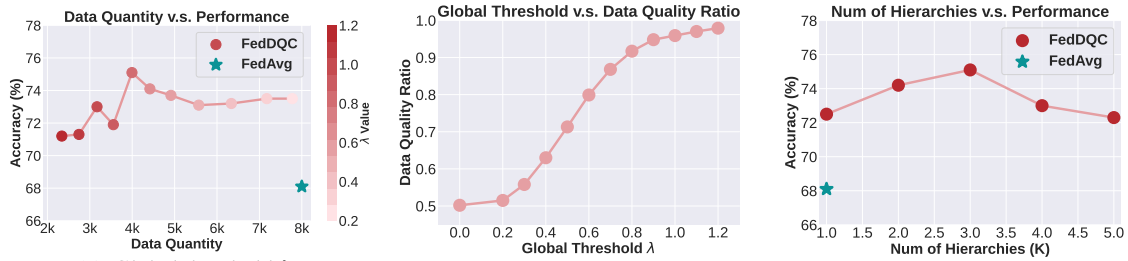
To demonstrate the close integration of IRA scores with hierarchical training, we compared 3 training sequences: random, ascending, and descending; across 4 domain-specific datasets in the IID setting, as shown in Table 2. The experiments reveal that: 1) IRA's relationship with easy-to-hard hierarchical training is mutually reinforcing, with descending sequence training significantly improving IRA-based data selection across all datasets. Notably, IRA consistently outperforms other quality evaluation metrics, regardless of the training sequence. 2) Other quality metrics do not consistently benefit from hierarchical training, highlighting their incompatibility with this approach.

6.4.2 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 4. 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, greatly enhancing data quality control.

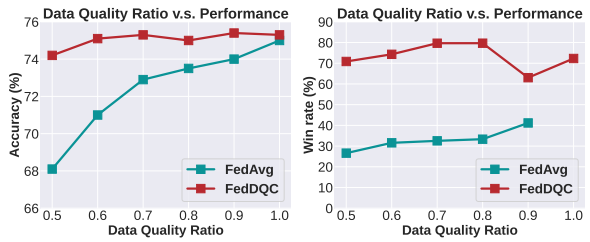
6.4.3 Data quality impact analysis

To examine how data quality impacts training, we quantify the dataset's overall quality as the ratio of aligned data to total data, and conduct experiments with varying data quality ratios (0.5 to 1.0) across four domain-specific datasets in the IID setting. For FiQA, we use win rates compared to the original dataset trained with FedAvg, so we exclude the 1.0 quality ratio for FedAvg. Key observations from Fig 7 include: 1) FedAvg performance decreases as the data quality ratio drops, showing the significant impact of low-quality data on training. 2) FedDQC outperforms FedAvg in all quality ratio settings, demonstrating the robustness of its data quality control. 3) Even with a quality ratio of 1.0 (no synthetic low-quality data), FedDQC consistently outperforms FedAvg, indicating its effectiveness in enhancing training performance, even in non-synthetic datasets.

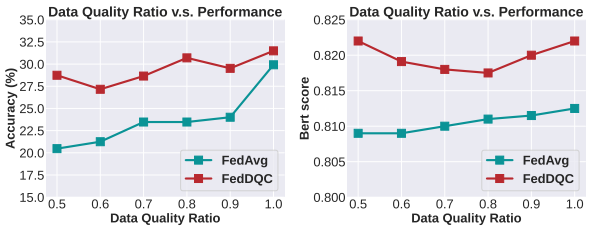


(a) Global threshold λ (b) Global thre. v.s. Data quality ratio (c) Number of hierarchies K

Figure 6: 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.



(a) PubMedQA (b) FiQA



(c) AQUA_RAT (d) Mol-Instructions

Figure 7: Comparison of FedAvg and FedDQC in various data quality ratios. (a)-(d) show the performance under different data quality ratio on various datasets. FedDQC is consistently higher than FedAvg in all data quality ratio on four datasets.

6.4.4 Convergence and Model Similarity analysis

To demonstrate FedDQC's impact on convergence and data heterogeneity, we compared local model similarity at round 50 between FedAvg and FedDQC in a quality NIID setting with 5 clients. Fig. 5 shows that, in FedAvg, model similarity is generally low due to data quality differences, particularly for client 5 with lower-quality data. In contrast, FedDQC's hierarchical training improves model similarity by filtering out low-quality data, reducing its negative impact, and enhancing aggregation. This results in a more stable global model, minimizing data heterogeneity and improving performance in heterogeneous settings.

6.5 Hyperparameter ablation

Global threshold To demonstrate the threshold robustness of FedDQC, we further examine the impact of the global threshold λ on the PubMedQA dataset with the IID setting. As shown in Fig. 6(a), the performance of FedDQC remains stable across varying λ values, indicating its insensitivity to the threshold. Even with changing data quantities, FedDQC consistently outperforms FedAvg, demonstrating its robustness. Additionally, as the threshold decreases, the data quality ratio in the selected data increases, see Figure 6(a) 6(c), but performance is more sensitive to the total data quantity than to data quality. This is evident from the asymmetric performance drop around 4k training data, where a decrease in data quantity results in a more pronounced performance decline.

Number of hierarchies Under the IID setting on PubMedQA, we tune the number of hierarchies in FedDQC $K \in \{1, 2, 3, 4, 5\}$. Figure 6(c) show that: 1) $K = 3$ is optimal; 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, too many hierarchies may reduce diversity, slightly hindering performance.

7 Conclusions

In this paper, we introduce FedDQC, a novel framework for data quality control in federated instruction-tuning of LLMs. FedDQC combines a new data quality assessment metric (IRA) with federated hierarchical training, where data quality is dynamically evaluated during training. Our experiments demonstrate that FedDQC adds minimal computational overhead while significantly improving model performance. The integration of IRA, adaptive scoring, and hierarchical training enhances both efficiency and robustness, making FedDQC a promising approach for effective controlled data quality in mixed-quality distributed scenario.

8 Limitations

A limitation of this study is that it assumes all local models share the same architecture, which is achievable when fine-tuning with the LoRA adapter. However, this approach may not be suitable for scenarios involving different local model architectures. Additionally, the study does not address the integration of data diversity into the design.

References

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

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

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

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

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

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

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

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

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. 2023. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*.

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

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

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

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

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

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

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

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

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. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

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

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

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

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

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

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

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5260–5271.

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

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

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

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450.

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

Robert F Ling. 1984. Residuals and influence in regression.

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

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

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

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

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

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

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

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

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

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

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*.

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

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

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping

and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293.

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

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

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

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

Lei Wang, Jieming Bian, and Jie Xu. 2024b. 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.

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

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

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

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

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

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

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

Jingyi Xu, Zihan Chen, Tony QS Quek, and Kai Fong Ernest Chong. 2022. 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.

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

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

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

Rui Ye, Rui Ge, Yuchi Fengting, Jingyi Chai, Yanfeng Wang, and Siheng Chen. 2024a. Leveraging unstructured text data for federated instruction tuning of large language models. *arXiv preprint arXiv:2409.07136*.

Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Yaxin Du, Yang Liu, Yanfeng Wang, and Siheng Chen. 2024b. Fedllm-bench: Realistic benchmarks for federated learning of large language models. *arXiv preprint arXiv:2406.04845*.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024c. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6137–6147.

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

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

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

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

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

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

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.

PubMedQA PubMedQA[1] (Jin et al., 2019) is a multiple-choice question-answering dataset optimized for medical reasoning. In this paper we utilize the version sourced from PMC-LLama (Wu et al., 2024). It features enhanced QA pairs with structured explanations derived from ChatGPT (Roumeliotis and Tselikas, 2023), facilitating in-depth medical analysis. PubMedQA dataset consists of 211.3k training samples.

FiQA FiQA dataset[2] is a subset from FinGPT (Yang et al., 2023), which consists 17.1k financial open question-answers. We split out 200 samples for evaluation and adopted the MT-Bench instruction template (see Table 10) to call ChatGPT (Roumeliotis and Tselikas, 2023) API (gpt-4-1106-preview). For the evaluation metric, we utilize the win rate to demonstrate the data quality ratio: $win_rate = win_counts/(win_counts + lose_counts)$.

AQUA_RAT The AQUA-RAT (Ling et al., 2017) 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.

Mol-Instructions The Mol-Instructions (Fang et al., 2023) 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 (Q&A) 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-Q&A task with 37k training set and 1k test set. For evaluation, the BertSocre (Zhang et al., 2019), an automatic evaluation metric for text generation, is applied on a predefined test set of size 200.

Fed-WildChat Fed-WildChat is a key component of the FedLLM-Bench (Ye et al., 2024b), a benchmark designed for evaluating FL methods in the context of LLMs. This dataset specifically focuses on multi-turn chat instruction tuning, providing a realistic representation of user-chatbot interactions. Fed-WildChat (Zhao et al., 2024) is derived from a collection of conversations between humans and ChatGPT, WildChat, featuring a diverse array of interactions. It comprises data from 100 clients, totaling approximately 53,000 samples. This dataset is structured to reflect real-world scenarios by partitioning the data based on user IP addresses, ensuring that each client has a substantial number of samples (at least 200) for effective training and evaluation

A.2 FedDQC algorithm

To control the data quality for training, two steps need to be conducted: data selection with data quality assessment, and training process with high-quality data. 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 2.

A.3 Baselines

Comparisons with current methods Compared to NUGGETS (Li et al., 2023b) and AlpaGasus (Chen et al., 2023c), which utilize an external model for quality evaluation, FedDQC evaluates the data on the client side and preserves local data privacy. Unlike DataInf (Kwon et al., 2023) and NUGGETS (Li et al., 2023b), 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 (Li et al., 2023a), FedDQC

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

Table 4: Dataset information and evaluation metrics

Dataset	Evaluation metrics	Domain	$\#samples$	$\hat{L}_{inst.}$	$\hat{L}_{Resp.}$
PubMedQA (Jin et al., 2019)	Acc	medical	211 k	471.1	71.4
FiQA (Yang et al., 2023)	Win rate	financial	17.1 k	42.1	255.7
AQUA-RAT (Ling et al., 2017)	Acc	math	97.5 k	77.4	105.7
Mol-Instructions (Fang et al., 2023)	Bert score	molecular	38 k	110.5	107.8
Alpaca-GPT4 (Peng et al., 2023)	-	general	52 k	21	163

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
for $k = 1$ to K **do**
  // Scoring Stage:
  for $n = 1$ to N **do**
    $\mathcal{D}_n = \mathcal{D}_n \setminus \mathcal{H}_{n(k-1)}$ ▷ Remove the trained data from $\mathcal{D}_n$
    $\mathcal{S}_n = \{s_i : s_i = f_{IRA}((q^i, a^i) \in \mathcal{D}_n, \theta^{(R/K)*(k-1)})\}$ ▷ 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
  end for
  // Training Stage:
  for $r = (R/K) * (k-1)$ to $(R/K) * k - 1$ **do**
    for $n = 1$ to N **do**
      Sort $\mathcal{D}'_n$ by quality scores s_i in descending order
      Split sorted $\mathcal{D}'_n$ into hierarchies $\mathcal{H}_{nk}, \ldots, \mathcal{H}_{nK}$ with equal size $floor(|\mathcal{D}'_n|/(K-k)$
      ▷ Split local dataset to hierarchies
      Local update θ^r_n with $\mathcal{H}_{nk}$ ▷ Local easy-to-hard hierarchical training
    end for
    $\theta^{r+1} = \sum_{n=1}^{N} w_n \theta^{r,t}_n$ ▷ Aggregate local models to update global model θ^r
    Distribute global model θ^{r+1} to each client n
  end for
end for
Return: Global model θ^R

does not require extra dataset adaptation training, thus, is computation effective.

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 (Wenzek et al., 2019).

DataInf Influence functions, a gradient-based scoring method, rely on the model's performance on a validation set. DataInf, as introduced by (Kwon et al., 2023), 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 $\mathrm{DataInf}(x_j)_i = \nabla L(x_j; \theta^\star) H^{-1}_{\theta^\star} \nabla L(x_i; \theta^\star)$ 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

Table 5: 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

$\text{IFD}_\theta(Q, A) = \frac{s_\theta(A|Q)}{s_\theta(A)}$, where $s_\theta(A) = -\frac{1}{N}\sum_{i=1}^{N} logP(w_i^A|w_1^A, ..., w_{i-1}^A; \theta)$, $s_\theta(A|Q) = -\frac{1}{N}\sum_{i=1}^{N} logP(w_i^A|Q, w_1^A, ..., w_{i-1}^A; \theta)$. 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 Experimental complements

A.4.1 Training setting

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 5 clients with around $8k$ training data in total for Fed-WildChat dataset. We randomly sample 2 clients for each round with 10 local steps using AdamW (Loshchilov and Hutter, 2017) 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 (Taori et al., 2023) to format the instruction, as shown in Appendix A.5.

A.4.2 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 5.

A.4.3 IRA metric analysis

Here, simplicity and complexity refer to learning difficulty. We analyzed the correlation between IRA scores and gradient magnitudes, finding that higher IRA scores correspond to smaller gradients, which indicate easier learning.

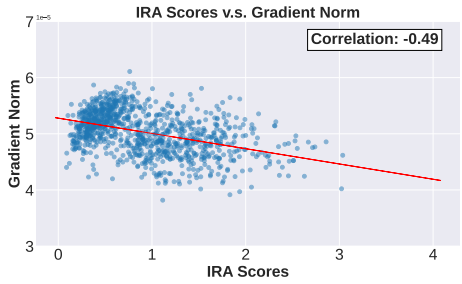


Figure 8: IRA score v.s. Gradient Norm

A.4.4 More FL settings.

We compare more federated settings with the number of clients equal to 5 and 20. See Table 7.

A.4.5 Other types of low-quality data

We have supplemented the experiments by adding comparisons with other baselines under different bad data construction scenarios, as well as mixed types of bad datasets. All datasets have 50% data corrupted.

The Table 6 below shows the proportion of high-quality samples globally before and after data selection, referred to as the data quality ratio. A ratio closer to 1 indicates more high-quality data. Our method, IRA, maintains a higher proportion of high-quality data after selection.

Table 6: Comparison of the Proportion of High - Quality Samples Before and After Data Selection: An Analysis of the Performance of Different Methods

	-	PPL	DataInf	IFD	NUGGESTS	IRA
Data quality ratio	0.5	0.8839	0.5003	0.700	0.6701	**0.9345**

Table 7: Comparison with other NIID settings and client numbers.

	client = 20		client = 5			
	IID	IID	NIID-0.1	NIID-1	NIID-5	NIID-10
oracle	0.741	0.750	0.737	0.743	0.747	0.758
mix-quality	0.691	0.681	0.662	0.655	0.664	0.685
selection	0.743	0.751	0.746	0.742	0.758	0.751

A.5 Prompt Template

A.6 Case Study

A.6.1 Examples of synthetic low-quality data

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 below.

Table 11, 12, 13, 14 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.6.2 Examples of scored data

Figures 9, 10, 11, 12 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 9 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: Performance comparison of FedAvg and various baseline methods under different bad data construction scenarios, with 50% data corruption across different types of data alteration strategies. The best-performing results are highlighted in bold.

Bad type	-	Swap	Delete	Cut	Substitute	Noisy	Mixture	Avg
FedAvg	0.757	0.681	0.691	0.730	0.720	0.734	0.700	0.709
FedAvg+PPL	-	0.703	**0.722**	0.694	0.689	0.727	0.644	0.696
FedAvg+DataInf	-	0.728	0.705	0.690	0.708	0.683	0.711	0.704
FedAvg+IFD	-	0.714	0.718	0.698	0.708	0.716	0.689	0.707
FedAvg+NUGGETS	-	0.708	0.102	0.301	0.269	0.722	0.477	0.429
FedDQC	0.750	**0.751**	0.710	**0.741**	**0.739**	**0.737**	**0.731**	**0.734**

Table 9: 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:

Table 10: 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]

Table 11: 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 12: 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 13: 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.

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 9: 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 10: The examples of scored data in mixed-quality FiQA datasets 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 11: 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 12: 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.

Table 14: 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.