ENHANCING DATA QUALITY IN FEDERATED FINE-TUNING OF LARGE LANGUAGE MODELS

Wanru Zhao[1]*, **Yaxin Du**[2]*, **Nicholas D. Lane**[1], **Siheng Chen**[2,3], **Yanfeng Wang**[2,3]
[1] University of Cambridge, [2] Shanghai Jiao Tong University,
[3] Shanghai AI Laboratory

ABSTRACT

In the current landscape of large language model training, there is a significant reliance on public domain data, which is nearing exhaustion according to recent research. To further scale up, it is crucial to incorporate collaboration among multiple specialized and high-quality private domain data sources. However, the challenge of training models locally without sharing private data presents numerous obstacles in data quality control. To tackle this issue, we propose a data quality control pipeline for federated fine-tuning of large language models. This pipeline computes scores reflecting the quality of training data and determines a global threshold for a unified standard, aiming for improved global performance. Our experiments show that the proposed quality control pipeline facilitates the effectiveness and reliability of the model training, leading to better performance.

1 INTRODUCTION

As businesses, products, and services spring up around large language models (LLMs, which we define as having more than one billion parameters), recent work have shown that one could attain better performance by training on more high-quality data (Kaplan et al., 2020; Hoffmann et al., 2022; Zhou et al., 2023). However, Villalobos et al. (2022) estimate that even high-quality English language data will be exhausted by the year 2024. What should we do when we run out of public data?

One solution is to exploit vast private data from various institutions, including enterprises and user devices. To unlock the potential of private data, it is essential to address the following two significant issues. First, it is imperative to preserve the privacy of all participants to protect their interests (Mondschein & Monda, 2019). To address this, one could adopt federated learning (McMahan et al., 2017), a collaborative machine learning framework that trains a model across multiple clients with their local private data, without exchanging any raw data. Second, enhancing model training requires the control the data quality from each participant. Due to the inability to directly access private data, quality control for these data poses significant challenges, which is the focus of this work.

Previously, data quality control relied heavily on manual selection processes (Touvron et al., 2023b;a). This approach, while commonly used, presented significant challenges due to the high volume of data, leading to substantial costs. Recent advancements have seen the introduction of automated low-quality data filters (Computer, 2023), such as perplexity filters (Muennighoff et al., 2023) and deduplication filters (Lee et al., 2021). These automated methods are designed to reduce data volume and enhance training efficiency in centralized settings, while their effectiveness in data quality control within collaborative environments remains to be explored.

In our paper, we propose an automated data quality control pipeline for federated fine-tuning of LLMs, showcasing notable performance improvements in mixed-quality data environments. Specifically, we incorporate data valuation algorithms to serve as scoring functions, enabling fine-grained evaluation of individual training sample quality. Furthermore, we establish a unified data quality standard using a minimal set of anchor data, addressing the challenge of heterogeneity in data quality across federated institutions. Adopting this approach, we effectively eliminate low-quality data, thereby enhancing model performance and ensuring privacy preservation. By leveraging the collaboration of

*Equal contribution. Corresponding to: Wanru Zhao (wz341@cam.ac.uk)

multiple private domain data sources, it opens up new possibilities in the face of real-world public data exhaustion.

2 Motivation and Setup: How low-quality data affects the performance of collaborative training

In our paper, we identify two unique challenges for federated fine-tuning of LLMs in terms of data quality. 1) **Real low-quality data** Firstly, we aim to highlight three prevalent patterns of low-quality data observed in real-world corpora: cut, deletion and exchange. The cut category encompasses scenarios where content is truncated due to word limit constraints, deletion pertains to instances where critical terminologies are absent from the corpus, and exchange refers to examples containing entirely incorrect information. We provide specific examples of these categories in Appendix E. 2) **Quality heterogeneity** Quality heterogeneity refers to the variability in the quality of data collection across different clients in federated learning. Given that federated learning often encompasses a vast number of clients, each with varying capabilities in data synthesis, it is impractical to assume uniformity in data quality among all participants. Consequently, some clients may possess a higher proportion of low-quality data compared to others, highlighting the absence of a uniform standard for data sample quality across all participants. We provide two Non-IID settings in Appendix B.

Figure 1: The impact of low-quality data on the performance of federated fine-tuning of LLMs .

In our preliminary experiments, we consider the two factors above, adjust the proportion of low-quality data of composition of PMC-LLama (Wu et al., 2023) and Medalpaca-flashcards (Han et al., 2023) datasets in federated training, shown in Figure 1. Higher scores indicate better performance (for more details about the metrics, see Appendix C.2). The key observation is, the quality of the training data has a significant effect on the performance of collaborative training: low-quality data consistently lead to worse influence on all the metrics.

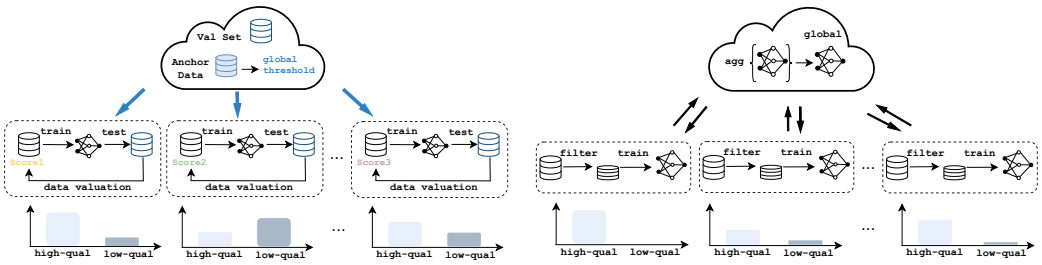
3 Proposed Workflow for Data Quality Control

3.1 Overview

Federated learning workflow iterates two steps on the client and server side: 1) client downloads global model $\boldsymbol{\theta}$ from the server and conducts training on local dataset $\mathcal{D}_k$ to obtain local model $\boldsymbol{\theta}_k$; 2) server receives local models $\{\boldsymbol{\theta}_k\}$ from clients and aggregates them to obtain global model $\boldsymbol{\theta} := \sum_k p_k \boldsymbol{\theta}_k$. In our workflow, each client conducts local training using its private, high-quality data. We have a public validation set located on both the clients and the server, which consists of commonly recognized, high-quality public data. As illustrated in Figure 2, the overall workflow consists of two phases designed to achieve data quality control in the federated training of LLMs.

Phase I. Pre-training: Data Quality Control Client receives initial global model θ^0 and public validation set $\mathcal{D}_{val}$ from server. Then clients compute each sample's quality score with scoring functions using the public validation set and global initial model. The server determines a global threshold of score, serving as a unified standard of data quality with only a very few amount of anchor data, and sends it to the clients.

Phase II. Federated Learning with High-Quality Data Each client then discards data samples that fall below the received global threshold, ensuring that only high-quality data verified by the unified standard is retained. Then clients utilize the filtered high-quality data sets $\mathcal{D}'_k$ (where $|\mathcal{D}'_k| \leq |\mathcal{D}_k|$) and global inital model θ^0 for federated learning.



(a) Phase I: Pre-training: Data Quality Control

(b) Phase II. FL with High-Quality Data

Figure 2: Overall workflow diagram consists of two phases: 1) Phase I: client-side compute each sample's quality score with scoring functions using the public validation set and global model, then server-side aggregates the scores, giving a global threshold by anchor data 2) Phase II: clients filter data according to the global threshold and starts federated learning on selected high-quality data.

3.2 Local Data Scoring and Quality Control

For each client, we have designed a data scoring step in which we calculate the score for each training data sample to assess its contribution to model prediction. Specifically, considering an ordered training set of examples $\mathcal{D}_k = \{z_1, \ldots, z_n\}$ and a model θ, we define the scoring function $S(z_i, \theta)$ for each data sample. Those with low scores are filtered out in subsequent steps. Following the inherent principle, we incorporate the concepts of existing measurements into our scoring method, which includes probability-based scoring and gradient-based scoring.

Perplexity: As a probability-based scoring method, perplexity is defined as the exponentiated average negative log-likelihood of a sequence. For a tokenized sequence $X = (x_0, x_1, \ldots, x_t)$, the perplexity scoring function of X is, $\mathrm{PPL}(X) = \exp\left\{-\sum_i^t \log p_\theta\left(x_i \mid x_{<i}\right)/t\right\}$, wherelog $p_\theta\left(x_i \mid x_{<i}\right)$ is the log-likelihood of the ith token conditioned on the preceding tokens $x_{<i}$ according to our model. It is commonly used as an evaluation of the model's ability to predict uniformly among the set of specified tokens in a corpus. Here it can be used as the scoring function for each individual data sample.

Conditional Probability: This is another probability-based scoring method. Each word in Question (Q) and Answer (A) is denoted as x_i^Q and x_i^A respectively. Conditional Probability on following instruction of a given (Q, A) pairs by calculating the ratio between $Pro_\theta(A)$ and $Pro_\theta(A \mid Q)$: $\mathrm{Prob}(X) = \sum_i \log p_\theta(x_i|x_{<i})/t, \mathrm{ConProb}(Q, A) = \mathrm{Prob}(A|Q)/\mathrm{Prob}(A)$. Conditional probability scores demonstrate how well the model aligns each response to the given corresponding question.

Influence Functions: This is a gradient-based scoring method based on the validation performance. The standard Hessian-based influence functions yield scores $\mathrm{DataInf}\left(x_j\right)_i = \nabla L\left(x_j; \theta^\star\right) H_{\theta^\star}^{-1} \nabla L\left(x_i; \theta^\star\right)$ for all $x_i \in \mathcal{D}_k$ and $x_j \in \mathcal{D}_{val}$, where $\theta^\star$ is the model trained on train set and $H_{\theta^\star}$ is the Hessian of the empirical loss. By fat, DataInf (Kwon et al., 2023) is the first computationally efficient influence approximation method that can be easily applied to LLM.

3.3 Global Standard with Anchor Data Scoring

On the server, we select only a few amount of data (10 samples in our paper) as our anchor data and use the aforementioned scoring method to calculate the average score of these 10 data points as the global threshold. This establishes a unified standard for division between low- and high-quality data for heterogeneous clients, allowing for the further filtering of local data.

4 Experiments

4.1 Experiment Setup

Tasks and Datasets We conduct our experiments on the question-answering (QA) task, sampled from the following two datasets. **1) PMC-LLama** (Wu et al., 2023) contributes a large-

scale, comprehensive dataset for instruction tuning. This dataset encompasses medical question-answering, rationale for reasoning, and conversational dialogues, comprising a total of 202M tokens. **2) Medalpaca-flashcards** (Han et al., 2023) presents an innovative dataset consisting of over 514k entries, specifically crafted to fine-tune LLMs for effective medical applications.

Models Our pre-trained base model is the LLama2-7b model (Touvron et al., 2023b), shown to perform well across many domains. We adapt commonly-used parameter-efficient fine-tuning technique, Low-Rank Adaptation (LoRA), which involves freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture, thereby achieving fine-tuning without incurring any additional inference latency (Hu et al., 2021).

Evaluation metrics In this study, we classify three evaluation metrics across five benchmark tests commonly utilized in contemporary medical LLMs. (Singhal et al., 2023; Wu et al., 2023; Lee et al., 2020; Luo et al., 2023) The evaluations focus on two main aspects: 1) **question-answering capabilities**, assessed by GPT-4 (OpenAI et al., 2023) and Openassistant (Köpf et al., 2023) scoring within the Medalpaca-flashcards test set. The former metric gauges the consistency of the generated responses with the ground truth, while the latter simulates the human reward of the answers. 2) **Knowledge acquisition ability**, measured by average accuracy of responses to multiple-choice questions in the MMLU (Hendrycks et al., 2021a), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and USMLE (Jin et al., 2021) datasets.

4.2 Experimental Results

Based on the low-quality dataset setup described in the Section 2, we build up our our data-quality control pipeline in centralized, federated IID and Non-IID setting (described in Appendix C.1). We implement both probability-based and gradient-based scoring methods to calculate scores for each training data, and set the unified scoring standard using corresponding scoring functions with anchor data. In comparison, we add a baseline, *Oracle*, which only trains on high-quality data. We also consider the commonly used in-context learning (ICL) as another scoring method, provided in Appendix C.4.

Table 1: Comparison of five data quality control methods across centralized, federated IID, and federated Non-IID settings with three evaluation metrics.

	Centralized ($n = 1$)			Federated ($n = 20$, IID)			Federated ($n = 20$, Non-IID)		
Evaluation Metric	GPT-4 Scoring	Open Assistant	Knowledge Avg	GPT-4 Scoring	Open Assistant	Knowledge Avg	GPT-4 Scoring	Open Assistant	Knowledge Avg
Low-qual Data	0.389	-0.89	0.398	0.411	-0.83	0.348	0.392	-0.87	0.396
Oracle	0.436	-0.63	0.442	0.455	-0.79	0.440	0.401	-0.75	0.417
PPL	**0.430**	**-0.84**	**0.399**	0.392	-0.92	**0.362**	**0.414**	-0.91	0.374
ConPro	**0.441**	**-0.75**	**0.445**	**0.420**	-0.86	**0.443**	**0.419**	-0.87	**0.444**
DataInf	0.381	**-0.74**	**0.412**	0.411	**-0.47**	**0.373**	**0.398**	**-0.66**	**0.423**
ICL	**0.438**	**-0.68**	**0.446**	**0.418**	**-0.86**	**0.437**	**0.435**	**-0.84**	**0.448**

The experimental results are presented in Table 1. Our findings include: **1)** Utilizing our current data quality control pipeline, our data scoring techniques consistently outperform the models trained on low-quality datasets in both centralized and federated settings. **2)** When employing *ConPro* and *ICL* as data scoring methods to select high-quality data, the performance of the global model can surpass that of models trained with the oracle set. **3)** In the Non-IID setting, the global model trained with our quality-controlled data demonstrates excellent performance. These experimental results highlight not only the efficacy of the data scoring methods but also the effectiveness of a unified quality threshold determined by anchor data (for more details, see Appendix D).

Conclusion and Future Work

In this paper, we establish a data quality control pipeline for federated fine-tuning of LLMs, avoiding directly sharing any private data. Preliminary experiments show that the selected high-quality data ensures an effective and reliable learning process, leading to improved model performance. To further safeguard privacy, we can seamlessly integrate differential privacy or a secure aggregation component to prevent reconstruction attacks, enhancing the security of our framework.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Together Computer. Redpajama: an open dataset for training large language models, 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Shizhe Diao, Rui Pan, Hanze Dong, Ka Shun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2306.12420*, 2023.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

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)*, 2021b.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL `https://doi.org/10.48550/arXiv.2203.15556`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL `https://arxiv.org/abs/2106.09685`.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

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.

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. *CoRR*, abs/2001.08361, 2020. URL `https://arxiv.org/abs/2001.08361`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

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

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

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*. PMLR, 2017.

Christopher F. Mondschein and Cosimo Monda. *The EU's General Data Protection Regulation (GDPR) in a Research Context*, pp. 55–71. Springer International Publishing, Cham, 2019. ISBN 978-3-319-99713-1. doi: 10.1007/978-3-319-99713-1_5. URL https://doi.org/10.1007/978-3-319-99713-1_5.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won

Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pp. 248–260. PMLR, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

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.

Yujun Shi, Jian Liang, Wenqing Zhang, Vincent YF Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. *arXiv preprint arXiv:2210.00226*, 2022.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.

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

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 2022.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=BkluqlSFDS`.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020b.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.

Wanru Zhao, Royson Lee, Yihong Chen, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

A Related Work

Large Language Models The remarkable achievements of large language models (LLM) have recently impacted the field of natural language processing. OpenAI's GPT-4 (OpenAI et al., 2023), for instance, has demonstrated exceptional capabilities in various generative tasks including question answering. LLaMA (Touvron et al., 2023a;b), an open-source large language modelwith 7 to 65 billion parameters, offers an alternative platform for research. These developments have sparked interest in adapting LLMs for medical applications. Yet, most medical models are fine-tuned based on LLaMA on a small medical corpus, resulting in a deficiency of comprehensive medical knowledge integration. There has been recent efforts on training LLMs for medical domains, for example, BioBert (Lee et al., 2020), BioMedGPT (Luo et al., 2023), PMC-LLama (Wu et al., 2023). These domain-specific LLMs have been exclusively trained on medical corpora. However, there has been a lack of collaborative or federated training work in medical LLMs.

Data Valuation and Attribution The seminal work on data valuation/attribution of Koh & Liang (2017) proposes attribution via approximate influence functions. It identifies training samples most responsible for a given prediction by estimating the effect of removing or slightly modifying a single training sample. In a related approach, TracIn (Pruthi et al., 2020) estimates the influence of each sample in training set on the test example by measuring the change in loss from gradient updates of mini-batches. Another related line of work has utilized Shapley values (Lundberg & Lee, 2017) to ascribe value to data, but Shapley values often require exponential time to compute.

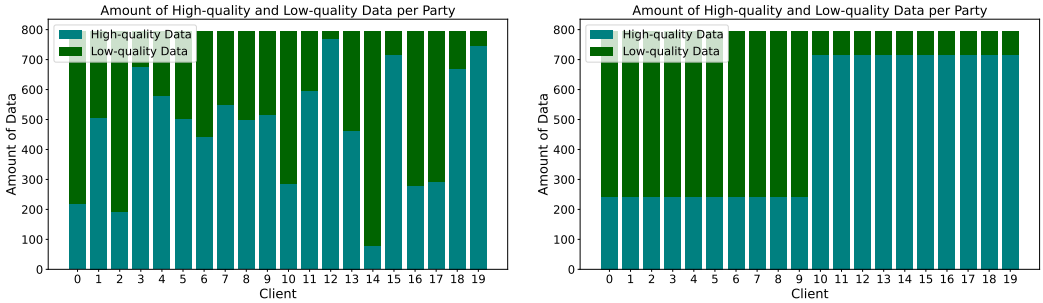
Federated Learning Federated Learning has garnered significant attention as a distributed machine learning paradigm. It shifts the traditional model training process by sharing model parameters instead of raw data. With Federated Averaging (FedAvg) (McMahan et al., 2017), participating clients train models using their own private datasets locally, and the updated model parameters are aggregated on the server. This preserves the privacy of the underlying data while collectively benefiting from the knowledge gained during the training process (Konečnỳ et al., 2016). Recent work (Zhao et al., 2024) proposes a federated parameter-efficient fine-tuning paradigm for large language models, demonstrating not only its advantages in data- and parameter-efficiency but also in better generalization and stability. Despite abundant research made on federated fine-tuning LLMs, data quality control in federated manner remains under-explored.

B Heterogeneity Settings

To model real-world scenario, we designed two heterogeneous settings: NIID-1 and NIID-2. NIID-1 replicates a typical scenario in federated learning classification tasks (Yurochkin et al., 2019; Wang et al., 2020a;b; Li et al., 2021; Shi et al., 2022), where the distribution of low-quality data among clients follows a Dirichlet distribution with parameter $\beta = 1$, while ensuring that the volume of data processed by each client remains equal. In contrast, NIID-2 addresses a skewed classification task scenario within FL (McMahan et al., 2017; Li et al., 2020), assigning 70% of low-quality data to half of the clients and 90% to the other half, yet maintaining an equal size of training data across all clients. The distributions for these settings are illustrated in Figure 3. Table2 shows the low-quality data traing and data quality control federated NIID-2 setting.

Table 2: Comparison of four data quality control methods on federated NIID-2 settings with three evaluation metrics.

Setting		GPT-4 Scoring	OpenAssistant	KnowledgeAvg
$n = 20$, NIID-2	Federated Low-qual Data	0.399	-0.94	0.399
	Oracle	0.418	-0.86	0.413
	ConPro	**0.427**	**-0.78**	**0.414**
	PPL	0.389	**-0.87**	0.199
	DataInf	0.381	-0.98	**0.424**
	ICL	0.341	**-0.85**	**0.426**



(a) NIID-1

(b) NIID-2

Figure 3: Data compromisation of high-quality and low-quality data with NIID-1 and NIID-2

C Experimental Details

C.1 Implementation Details

Training setups We use Hugging Face's transformers library (Wolf et al., 2020) and PEFT library (Mangrulkar et al., 2022) for loading pre-trained models and LoRA configurations. For our training and evaluation, we use PyTorch as the underlying auto-differentiation framework (Paszke et al., 2019). We use the AdamW optimizer (Loshchilov & Hutter, 2019; Kingma & Ba, 2015) for all experiments. All the experiments are conducted on the machines with the same hardware configuration using NVIDIA A40. In all experiments, we use 8 bit quantization, set $batch_size = 16$, and LoRA rank to 16. For the federated setting, we consider 300 communication rounds, $n = 20$ clients, 10 local steps of model training (equal to 3 epoch for local training).

Training data We use 16k sample in total, with 8k samples randomly sampled from PMC-LLama (Wu et al., 2023) and Medalpaca-flashcards (Han et al., 2023) each. In the low-quality data synthetic processes, 3.2k samples (40% total data) are either polluted with cut (10% total data), delete (15% total data) and exchange (15% of total data) tricks. These 40% low-quality data together with the rest of high-quality data composites the low-quality data set. While the oracle data set only includes the rest of high-quality data set.

C.2 Evalution metrics

GPT-4 scoring 200 samples are randomly selected from the Medalpaca-flashcards dataset to serve as the test set. We evaluate the models that need to be compared on the test set to generate responses respectively. Then we use OpenAI GPT-4 model API to assign scores for their responses. Each response of is rated by the judge on a scale from 0 to 1, reflecting how well the answer aligns with the ground truth.

OpenAssistant OpenAssistant (Köpf et al., 2023) is a reward model built on the DeBERTa (He et al., 2020) architecture, trained with human feedback data from studies by Nakano et al. (2021); Stiennon et al. (2020); Bai et al. (2022). In this metric, 200 samples were randomly selected from the Medalpaca-flashcards dataset as the test set, with each assigned a score from -5 to 5 to reflect data quality.

Knowledge-based benchmarks For evaluation on multiple-choice questions, we use the USMLE (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019) and MMLU clinical topics (Hendrycks et al., 2021b) datasets. More details refer to Appendix C.4.

C.3 In Context Learning

In-context learning primarily denotes the capacity of LLMs to adapt to new information or tasks based solely on the context presented in their input, eliminating the necessity for explicit retraining or

fine-tuning for those specific tasks. Meanwhile, some work (Zhou et al., 2023) hypothesizes that most knowledge in LLM is learned from the pre-training stage, while instruction tuning stimulates the instruction-following ability. Therefore, in the realm of data selection, we assume pre-trained LLM is able to select obvious low-quality data by prompting pre-trained model with serveral examples. Sepicifally, incontext learning method provides several examples $(\hat{Q}, \hat{A})$ to a pre-trained model with respect to specific sample (Q, A). we form the input as $\overline{Q} = [(\hat{Q}, \hat{A}), Q]$, and take the in-context learning score as $l(A, p_\theta(A|\overline{Q}))$, where l is the training loss.

C.4 Knowledge-based benchmark

MMLU MMLU (Hendrycks et al., 2021b) provides a comprehensive suite of tests for assessing text models in multi-task contexts. We utilized the clinical topics from the MMLU test set as our testing ground, encompassing a diverse range of subjects including 265 questions on Clinical Knowledge, 100 on Medical Genetics, 135 on Anatomy, 272 on Professional Medicine, 144 on College Biology, and 173 on College Medicine, all formatted as multiple-choice questions.

MedMCQA MedMCQA (Pal et al., 2022) is a dataset comprised of multiple-choice questions derived from mock exams and past papers of two major Indian medical school entrance examinations, namely AIIMS and NEET-PG. The dataset is divided into two parts: the training split, which includes 182,822 questions, and the test split, consisting of 4,183 questions. Each question in the dataset is accompanied by four possible answers.

PubMedQA PubMedQA (Jin et al., 2019) is a biomedical QA benchmark collected from PubMed abstracts. The PubMedQA task is designed to answer research questions with responses categorized as yes/no/maybe, effectively framing it as a multiple-choice question format. The dataset is divided into three subsets: 1,000 manually labeled question-answer pairs (denoted as PQA-L), 61,200 unlabeled pairs (PQA-U), and 211,300 pairs that have been artificially generated (PQA-A). Consistent with previous studies (Diao et al., 2023; Singhal et al., 2023), we employ the PQA-L subset as the test set for evaluating the model's performance.

USMLE USMLE (Jin et al., 2021) consists of multiple-choice questions (with 4 choices per question) that are based on the United States Medical Licensing Exams. This dataset has been compiled from questions used in professional medical board examinations and is unique in its multilingual composition, including English, Simplified Chinese, and Traditional Chinese versions. It contains 12,724 questions in English, 34,251 in Simplified Chinese, and 14,123 in Traditional Chinese. For our purposes, we focus on the English component of the dataset, which is further divided into 10,178 questions for the training set, 1,273 for the validation set, and 1,273 for the test set, adhering to the official distribution of the dataset.

D Ablation study of unified scoring with anchor data

We perform a comparative analysis focusing on the number of selected data and the proportion of low-quality data. This comparison contrasts the approach of selection based on a unified scoring from anchor data, as depicted in the fourth bar, with methods that determine the threshold by acknowledging the overall proportion of low-quality data. In Figure 4, the method represented by the second bar involves each client selecting high-quality data by sorting each sample's scores and eliminating the lowest-scored samples, informed by the proportion of low-quality data within the entire training set of FL. This approach, which does not need a score threshold and relies solely on the known proportion of low-quality data, proves to be impractical and less effective in settings with heterogeneous data quality due to varying data quality compositions among clients. The third bar illustrates the outcomes of implementing a global score threshold aligned with the exact proportion of global low-quality data. This method necessitates the server's aggregation of all scores from clients within the FL framework and knowing the proportion of low-quality data, potentially leading to privacy concerns. The results indicate that selection by anchor score consistently yields a lower proportion of low-quality data compared to the other two methodologies.

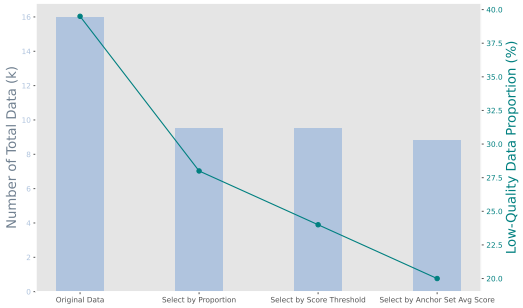


Figure 4: Number of selected data and the proportion of low-quality data across different selection principles (select by proportion, score threshold, or anchor set score) in federated NIID1 setting, employing the ConPro score.

E EXAMPLES FOR LOW- AND HIGH- QUALITY DATA

To model real-world scenario, we designed three types of low-quality data generation rules.

Cut The cut type data simulates scenarios where an answer exceeds the word limit, resulting in the truncation of the last few words. Practically, we establish a threshold of 100 tokens, retaining only the initial tokens up to this limit, as demonstrated in Table 5.

Deletion Deletion type of data refers to missing important terminologies. In practice, we delete part of the tokens in the answer to simulate the missing terminology or some nouns, shown in Table 4.

Exchange Exchange type low-quality data is characterized by the complete irrelevance between questions and their corresponding answers. In this approach, we swap answers among question-answer pairs to illustrate this form of low-quality data. Table 3 demonstrates examples of this exchange, where the first two samples involve swapping answers within the Medalpaca-flashcards dataset (Han et al., 2023). The remaining two samples, drawn from PMC-Llama (Wu et al., 2023), undergo a process where their rationales for answer choices are exchanged, and a different, incorrect choice is selected to ensure a clear distinction in the option outputs.

Index	Question	Original Answer	Answer
11136	'###Question: What is the name of the agent used to treat pulmonary hypertension, and how does it work in relation to the Endothelin-1 (ET-1) receptor?'	'###Answer: The agent used to treat pulmonary hypertension is called Bosentan, and it works by blocking the Endothelin-1 (ET-1) receptor.'	'###Answer: In multiple myeloma, neoplastic plasma cells produce IL-6 which stimulates osteoclast activating factor.'
13747	'###Question: What specific substance do neoplastic plasma cells produce in multiple myeloma that stimulates osteoclast activating factor?'	'###Answer: In multiple myeloma, neoplastic plasma cells produce IL-6 which stimulates osteoclast activating factor.'	'###Answer: Permethrin is the anti-scabies/louse therapy that causes neuronal membrane depolarization.'
4888	'###Question: Iodine deficiency control programme ?\n###Options:\nA. Health education\nB. Water testing\nC. Foification of salt\nD. None\n'	"###Rationale: Ans. is 'c' i.e., Foification of salt Iodized salt (salt foified with iodine) is most economical, convenient and effective means of mass prophylaxis in endemic area.\n\n###Answer: OPTION C IS CORRECT."	'###Rationale: Fracture neck of talus results from forced dorsiflexion of the ankle. Typically this injury is sustained in an aircraft crash where the rubber bar is driven forcibly against the middle of the sole of the foot (ATOR&;s fracture), resulting in forced dorsiflexion of the ankle, the neck being a weak area, gives way. Reference - Essential ohopaedics-Maheshwari -5th edn pg no 166.\n\n###Answer: OPTION B IS CORRECT.'
4066	'###Question: Ator fracture is -\n###Options:\nA. Fracture neck of talus\nB. Fracture scaphoid\nC. Fracture calcaneum\nD. Fracture 5th metatarsal\n'	'###Rationale: Fracture neck of talus results from forced dorsiflexion of the ankle. Typically this injury is sustained in an aircraft crash where the rubber bar is driven forcibly against the middle of the sole of the foot (ATOR&;s fracture), resulting in forced dorsiflexion of the ankle, the neck being a weak area, gives way. Reference - Essential ohopaedics-Maheshwari -5th edn pg no 166.\n\n###Answer: OPTION A IS CORRECT.'	"###Rationale: Ans. is 'c' i.e., Foification of salt Iodized salt (salt foified with iodine) is most economical, convenient and effective means of mass prophylaxis in endemic area.\n\n###Answer: OPTION D IS CORRECT."

Table 3: Synthetic low-quality examples of exchange type

Index	Question	Original Answer	Answer
12368	'###Question: What are some potential health risks associated with excess intake of Vitamin C?'	'###Answer: Excess intake of Vitamin C can lead to a number of potential health risks, including the formation of calcium oxalate kidney stones, a condition known as calcium oxalate nephrolithiasis. This is because Vitamin C is broken down in the body to oxalate, a compound that can combine with calcium to form kidney stones. Other potential risks associated with high doses of Vitamin C include gastrointestinal issues such as diarrhea, nausea, and abdominal cramps. It is important to consume Vitamin C in moderation and to follow the recommended daily intake levels to avoid these potential health risks.'	'###Answer: intake can lead a number of risks, the of calcium oxalate stones, a as calcium nephrolithiasis. because Vitamin C broken down in body to oxalate, form kidney Other risks with high of Vitamin include gastrointestinal issues as It to Vitamin C and to follow levels avoid these potential'
7297	'###Question: Does liver damage in the rat induce hepatocyte stem cells from biliary epithelial cells?\n###Context: When rat hepatocyte regeneration after partial hepatectomy is blocked by 2-acetylaminofluorene, a proliferation of biliary epithelia sends out ductules into the parenchyma. The ability of these neoductules to act as a significant progenitor compartment for hepatocytes is in dispute. This study aims to resolve this question by varying the amount of 2-acetylaminofluorene administered. Rats were fed 2-acetylaminofluorene fr 6 days before and up to 7 days after partial hepatectomy was performed at a dose of either 2.5 (low) or 5 (high) mg/kg(-1)/day(-1). The response was monitored by the immunohistochemical expression of intermediate filaments and cytochrome P450 enzymes. No regeneration by mature hepatocytes occurred with either dose, and new ductules expressed the biliary cytokeratins 7, 8, 18, and 19 and, in addition, vimentin. At the high dose, hepatocytic differentiation was infrequent, whereas apoptosis and intestinal differentiation were common. At the low dose, almost all ductules differentiated into hepatocytes within 14 days of hepatectomy.'	'Biliary epithelium is an effective and substantiative hepatocyte progenitor compartment under appropriate conditions.\n\n###Answer: yes'	'Biliary effective and substantiative hepatocyte under conditions.\n\n### Answer:'

Table 4: Synthetic low-quality examples of delete type

Index	Question	Original Answer	Answer
8002	'###Question: Which antibiotics are commonly prescribed to treat hepatic encephalopathy?'	"###Answer: The antibiotics that are typically used to treat hepatic encephalopathy are rifaximin and neomycin. Hepatic encephalopathy is a neurological disorder that occurs as a result of liver disease. It is characterized by a range of symptoms such as confusion, disorientation, and impaired motor function. Antibiotics such as rifaximin and neomycin are used to reduce the levels of ammonia in the body, which can contribute to the development of hepatic encephalopathy. These antibiotics work by reducing the number of bacteria in the gut that produce ammonia. Rifaximin is preferred due to its lower risk of side effects compared to neomycin. However, the choice of antibiotic may depend on the individual patient's medical history and the severity of their condition."	'###Answer: The antibiotics that are typically used to treat hepatic encephalopathy are rifaximin and neomycin. Hepatic encephalopathy is a neurological disorder that occurs as a result of liver disease. It is characterized by a range of symptoms such as confusion, disorientation, and impaired motor function. Antibiotics such as rifaximin and neomycin are used to reduce the levels of ammonia in the body, which can contribute to the development of hepatic encephalopathy. These antibiotics work by reducing the number of bacteria in the gut that produce ammonia. Rifaximin is preferred due to its lower risk of side effects compared to neomycin.'
362	'###Question: Which of the following statement is TRUE about NPU (net protein utilization), used in protein quality assessment?\n###Options:\nA. It is the ratio of energy from protein to total energy in diet\nB. It is the ratio between nitrogen retained by the body and total nitrogen intake multiplied by 100\nC. It is the amount of one amino acid per gram of a protein divided by the amount of same amino acid per gram of egg protein\nD. If the NPU is high the amount of protein requirement in diet is high\n'	"###Rationale: Net Protein Utilization (NPU) is the ratio between nitrogen retained by the body and total nitrogen intake multiplied by 100. It is a product of digestibility coefficient and biological value divided by 100. It gives a more complete expression of protein quality than the amino acid score. In calculating protein quality, 1 gram of protein is assumed to be equivalent to 6.25 g of N. Ref: Park's Textbook of Preventive and Social medicine, 19th Edition, Page 503.\n\n###Answer: OPTION B IS CORRECT."	"###Rationale: Net Protein Utilization (NPU) is the ratio between nitrogen retained by the body and total nitrogen intake multiplied by 100. It is a product of digestibility coefficient and biological value divided by 100. It gives a more complete expression of protein quality than the amino acid score. In calculating protein quality, 1 gram of protein is assumed to be equivalent to 6.25 g of N. Ref: Park's Textbook of"

Table 5: Synthetic low-quality examples of cut type