# OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning


Rui Ye
Shanghai Jiao Tong University
Shanghai, China
yr991129@sjtu.edu.cn

Wenhao Wang
Zhejiang University
Zhejiang, China
12321254@zju.edu.cn

Jingyi Chai
Shanghai Jiao Tong University
Shanghai, China
chaijingyi@sjtu.edu.cn

Dihan Li
University of Southern California
Los Angeles, USA
dihanli@usc.edu

Zexi Li
Zhejiang University
Zhejiang, China
zexi.li@zju.edu.cn

Yinda Xu
Shanghai Jiao Tong University
Shanghai, China
yinda_xu@sjtu.edu.cn

Yaxin Du
Shanghai Jiao Tong University
Shanghai, China
dorothydu@sjtu.edu.cn

Yanfeng Wang
Shanghai Jiao Tong University,
Shanghai AI Laboratory
Shanghai, China
wangyanfeng@sjtu.edu.cn

Siheng Chen*
Shanghai Jiao Tong University,
Shanghai AI Laboratory
Shanghai, China
sihengc@sjtu.edu.cn


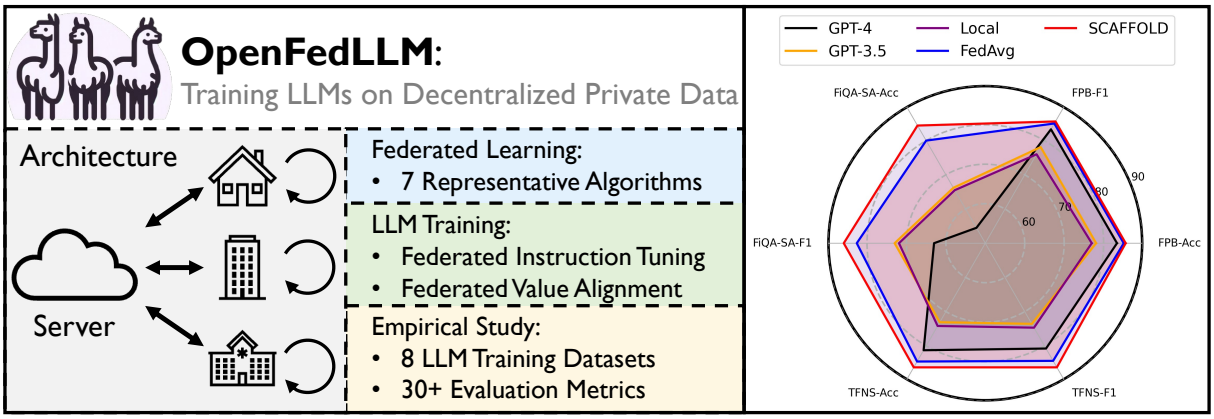


**Figure 1: Overview of our proposed `OpenFedLLM` framework and one example of experimental results. `OpenFedLLM` integrates 7 representative federated learning algorithms, federated instruction tuning, and federated value alignment and supports 8 training datasets and 30+ evaluation metrics. The experiments (right) showcase the results of federated instruction tuning on the financial domain, where we see that FL helps train a better LLM that can outperform GPT-4 and GPT-3.5.**


## Abstract

Trained on massive publicly available data, large language models (LLMs) have demonstrated tremendous success across various fields. While more data contributes to better performance, a disconcerting reality is that high-quality public data will be exhausted in a few years. In this paper, we offer a potential next step for contemporary LLMs: collaborative and privacy-preserving LLM training on the underutilized distributed private data via federated learning (FL), where multiple data owners collaboratively train a shared model without transmitting raw data. To achieve this, we build a concise, integrated, and research-friendly framework/codebase, named `OpenFedLLM`. It covers federated instruction tuning for enhancing instruction-following capability, federated value alignment for aligning with human values, and 7 representative FL algorithms. Besides, `OpenFedLLM` supports training on diverse domains, where we cover 8 training datasets; and provides comprehensive evaluations, where we cover 30+ evaluation metrics. Through extensive experiments, we observe that all FL algorithms outperform local training on training LLMs, demonstrating a clear performance improvement across a variety of settings. Notably, in a financial



*Siheng Chen is the corresponding author.




*KDD '24, August 25–29, 2024, Barcelona, Spain.*



ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671582

benchmark, Llama2-7B fine-tuned by applying any FL algorithm can outperform GPT-4 by a significant margin, while the model obtained through individual training cannot, demonstrating strong motivation for clients to participate in FL. The code is available at https://github.com/rui-ye/OpenFedLLM. The full version of our paper is available at https://arxiv.org/pdf/2402.06954.

## CCS Concepts

• **Computing methodologies** → **Distributed computing methodologies**; • **Security and privacy** → **Human and societal aspects of security and privacy**; • **Information systems** → *Information systems applications*.

## Keywords

Large Language Models, Federated Learning, Instruction Tuning, Value Alignment

**ACM Reference Format:**
Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. 2024. 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 (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671582

## 1 Introduction

Trained on massive public data, large languages models (LLMs) [1, 11, 29, 55, 56, 72] have demonstrated tremendous success across a broad spectrum of fields in recent years [7, 25, 61, 62, 80, 82]. Nevertheless, an issue of significant concern has emerged amidst this proliferation of LLMs: it has been estimated that high-quality public data will be exhausted before 2026 [74]. The scarcity of data can also be discerned from a current trend where more researchers tend to train data-hungry LLMs by combining existing datasets [77] or using model-generated datasets [78, 86], rather than collecting and generating new datasets. These observations indicate that the development of current LLMs could potentially come to a bottleneck since the commonly acknowledged scaling laws show that more data usually leads to better performance [31].

Meanwhile, an abundance of high-quality data is distributed across diverse parties but remains underutilized, which cannot be publicly shared due to issues such as privacy (e.g., medical [70] and financial [85] data) or physical constraints (e.g., lacking network connections). As a representative case, trained on large amounts of private financial data (over a span of 40 years), BloomberGPT [85] demonstrates exceptional performance in finance, indicating the value of high-quality private data. However, the challenge lies in the fact that not every party possesses sufficient data to train a well-performed and data-hungry LLM individually.

Considering the limitations of public data and the high utility yet potential scarcity of one's private data, it is critical to enable the advancement of modern LLMs through **collaborative training on decentralized private data without direct data sharing**.

In this paper, we comprehensively explore the potential of training LLMs on the underutilized decentralized private data via federated learning (FL) [53], a privacy-preserving training paradigm where multiple parties collaboratively train a model under the coordination of a central server [30]. Specifically, starting from an off-the-shelf base LLM that has been pre-trained on a large corpus, we aim to train/fine-tune the LLM to achieve interested functionalities via FL, which consists of four iterative steps: global model downloading, local model training, local model uploading, and global model aggregating. Here, we focus on two critical and representative procedures in the training of contemporary LLMs: instruction tuning [48, 56, 86, 95] and value alignment [2, 28, 34, 56], positioning as two applications in collaborative and privacy-preserving training of LLMs on decentralized private data.

In federated instruction tuning (FedIT), we adopt the conventional supervised fine-tuning (SFT) method [56] during local training for each client, where each data sample is an instruction-response pair, and the LLM is trained to predict the response given the instruction. With FedIT, the LLM can be trained to follow humans' diverse instructions, which is achieved by unifying massive clients to join the FL system. However, human values are not well included during FedIT, resulting in some imperfections, such as failing to ensure safe responses from the LLMs. Therefore, a subsequent stage for value alignment is commonly required. In federated value alignment (FedVA), we adopt one of the most stable training methods to date, direct preference optimization (DPO) [58], during local training. During this process, each instruction is accompanied by one preferred response and another dispreferred response, where the LLM is trained to align with the preference and keep away from the dispreference. With FedVA, human values can be injected into the LLMs, which can be strengthened by involving a large number of clients to cover diverse human values.

To enable an exhaustive exploration, we build a concise, integrated, and research-friendly framework named `OpenFedLLM`, where the users can easily focus on either FL or LLMs without much background knowledge of the other field (LLMs or FL); see Figure 1 for an overview. In `OpenFedLLM`, we **(1)** implement diverse critical features, covering federated instruction tuning, federated value alignment, multiple representative FL baselines (i.e., 7), diverse training datasets (i.e., 8) and evaluation metrics (i.e., 30+), and more; **(2)** make huge efforts to decouple the implementation of FL and LLM training, reducing the engineering cost of both two communities and thus encouraging their joint future contributions. Besides, we apply quantization and parameter-efficient fine-tuning [24] techniques together with memory-saving strategies [6], making the training executable on one single consumer GPU (e.g., NVIDIA 3090). It is worth noting that `OpenFedLLM` is the first framework that simultaneously integrates federated instruction tuning, federated value alignment, and diverse FL baselines, contributing to bridging the gap between these two communities.

Based on our `OpenFedLLM` framework, we provide a comprehensive empirical study on 7 baselines, 8 datasets, 30+ evaluations, and multiple configurations (e.g., in-domain collaboration and cross-domain collaboration), offering new insights and better understanding for future research. Through extensive experiments, we have several key observations. (1) FL can always bring benefits compared to individual training on the training of LLMs, offering strong motivation for organizations (especially those with limited data) to participate in FL to train better LLMs. (2) Training of LLMs via FL only requires one single GPU and takes $1-2$ hours per client for

100 communication rounds. (3) No FL algorithm can guarantee the best performance in all scenarios. (4) Under some specific domains, such as finance, that require domain-specific expert knowledge, FL on the corresponding dataset can even outperform GPT-4 [55] (the most excellent LLM to date) with an evident gap. Note that this is the first time in the literature showing that FL can outperform GPT-4 at any dimension.

Looking forward, we anticipate that others will build upon our `OpenFedLLM` framework for further explorations. (1) In FedLLM, new challenges and directions are emerging, such as heterogeneous preferences in FedVA, logically correct yet harmful attackers, and data management of decentralized private data, all of which call for future efforts. (2) Since currently no FL algorithm dominates in all scenarios, we expect to see new FL algorithms specifically tailored for LLMs training, serving as effective and pioneering representatives in FedLLM. (3) In this era of LLMs, we advocate future works in FL communities to implement their algorithms in our framework to examine their performance in such new application scenarios, making FL evolve with the recent trends.

Our contributions are as follows:

(1) We explore the complete pipeline for fine-tuning contemporary large language models on decentralized private data resources via federated learning, pointing out a promising development direction for LLMs.
(2) We propose an integrated and concise codebase/framework `OpenFedLLM`, covering applications of instruction tuning and value alignment, diverse FL baselines, training datasets, and evaluation datasets, which is research-friendly for both communities of LLMs and FL.
(3) We present a comprehensive empirical study based on our `OpenFedLLM`, showing FL methods consistently outperforms individual training (e.g., $\geq$ 12% improvement on MT-Bench on general dataset). We also offer new insights and point out research directions for future work.

## 2 Related Work

### 2.1 Large Language Models

Large language models (LLMs) such as GPT-3.5/4 [55, 56] and Llama2 [72] have demonstrated success in diverse domains [35, 37, 75, 85]. These contemporary LLMs are usually trained in three stages: (1) auto-regressive pre-training on large corpus such as C4 [59] and Pile [21], where the LLMs learn world's general knowledge [3, 64, 71]. (2) Instruction tuning on instruction-response pairs where the LLMs learn to follow instructions [81, 86, 95]. (3) Value alignment on human-annotated or AI-annotated preference dataset where humans' value is injected into the LLMs [1, 39, 56].

Currently, these steps are mostly conducted on publicly available data, which is either publicly released [48, 95] or AI-generated [9, 57, 69, 86]. However, it has been estimated that high-quality public data will exhaust before 2026 [74], indicating a forthcoming bottleneck of current LLMs since more data usually contributes to better performance [31]. Therefore, recently, there have been several attempts that train LLMs on privately-kept data [66, 72]. For example, trained on financial data spanning 40 years, BloombergGPT [85] has demonstrated strong performance in finance.

**Table 1: Comparisons among `OpenFedLLM` and other FL frameworks. IT: instruction tuning, VA: value alignment, $N_{FL}$: number of supported FL algorithms, $N_{TD}$: number of training datasets, $N_{EM}$: number of evaluation metrics.**

| Framework Name | IT | VA | $N_{FL}$ | $N_{TD}$ | $N_{EM}$ |
|---|---|---|---|---|---|
| FATE-LLM [17] | × | × | 1 | 1 | 4 |
| Shepherd [93] | ✓ | × | 1 | 1 | 1 |
| FederatedScope-LLM [36] | ✓ | × | 1 | 3 | 3 |
| **`OpenFedLLM` (ours)** | ✓ | ✓ | **7** | **8** | **30+** |

However, in the real world, the data amount of each party could be limited, while the union of massive parties' data could form a large database to train a powerful LLM [77]. Therefore, it becomes imperative to contemplate the forthcoming evolution of LLMs: collaborative training on distributed private data in a privacy-preserving way. Since pre-training often requires high compute resource [64] and is inapplicable with parameter-efficient tuning techniques such as LoRA [24], this paper focuses on the last two steps: instruction tuning and value alignment.

### 2.2 Federated Learning

Fortunately, federated learning (FL) [30] offers great potential to empower achieving privacy-preserving collaborative training. FL enables multiple parties (i.e., clients) to collaboratively train a shared global model without transmitting raw data under the coordination of a central server [53]. Typically, FL involves four steps: server-to-client global model broadcasting, local model training at the client, client-to-server local model uploading, and global model updating via aggregation at the server.

Since the vanilla FL method FedAvg [53] could only achieve moderate performance, especially under scenarios of data heterogeneity [23, 43], many FL algorithms are proposed to boost the performance of FL. (1) On the client side, there are methods that focus on enhancing consistency among local models and, therefore, boosting the performance of the aggregated model [32, 88]. FedProx [41] proposes to regularize the distance between local and global models. SCAFFOLD [33] introduces control variate to correct gradients of local models. (2) On the server side, there are methods that focus on refining the aggregation process and, therefore, improving the performance of global model [10, 42, 45]. FedAvgM [23] and FedOPT [60] introduce momentum for updating the global model. FedNova [76] and FedDisco [90] focus on modifying the weights for aggregating local models.

The performance of these methods has been verified mostly in the context of image classification and small models; however, their performance in current LLM training remains unclear. Therefore, in this paper, we are the first to explore their behaviors in the context of LLM training, providing new insights and searching for appropriate methods for federated LLM training.

### 2.3 FL and LLMs

Recently, there have been several preliminary works about FL and LLMs [26]. Some release a position paper while no empirical results are provided [4]. FATE-LLM [17] explores federated fine-tuning on LLMs, which is limited to conventional tasks (i.e., advertise

generation) rather than instruction tuning or value alignment. FederatedScope-LLM [36] and Shepherd [93] both explore FedIT. However, they are limited for the following three reasons. First, their empirical results are not sufficient enough as their training and evaluation datasets are relatively limited (e.g., Shepherd [93] is based on 1 training and 1 evaluation dataset). Second, none of them consider value alignment, which is imperative for releasing modern Chatbots [55]. Third, both of them are limited to FedAvg [53] as the only FL method while neglecting the diverse FL algorithms that have been shown to perform better depending on the tasks.

Unlike previous works, we provide the most comprehensive exploration of FL and contemporary LLMs to date. From the perspective of LLMs, we explore two critical steps in the current LLMs training paradigm, including instruction tuning and value alignment. From the perspective of FL, we explore 7 representative FL algorithms. Besides, we provide a comprehensive empirical study, covering 8 training datasets and over 30 evaluation metrics.

## 3 OpenFedLLM Framework

In this section, we first overview `OpenFedLLM`. Then, we introduce two critical procedures in `OpenFedLLM`: federated instruction tuning, which enhances instruction-following capability, and federated value alignment, which enhances alignment with human values.

### 3.1 Overview of OpenFedLLM

To make our framework compatible with standard FL protocols such as secure aggregation and differential privacy, our `OpenFedLLM` framework follows the same training process of conventional FL (i.e., FedAvg [53]). The overall process takes $T$ communication rounds, where each round $t$ consists of four key steps. (1) The server broadcasts the global model $\boldsymbol{\theta}^t$ to all available clients $\mathbb{S}^t$; (2) Each available client $k$ executes $\tau$ steps of SGD on its local dataset $\mathcal{D}_k$ starting from the global model $\boldsymbol{\theta}^t$, resulting a local model denoted as $\boldsymbol{\theta}_k^{(t,\tau)}$; (3) Each available client $k$ uploads the local model $\boldsymbol{\theta}_k^{(t,\tau)}$ to the server; (4) The server aggregates the local models and updates the global model for the next round: $\boldsymbol{\theta}^{t+1} := \sum_k^{\mathbb{S}^t} p_k \boldsymbol{\theta}_k^{(t,\tau)}$, where $p_k = \frac{|\mathcal{D}_k|}{\sum_i^{\mathbb{S}^t} |\mathcal{D}_i|}$ is the relative dataset size.

On one hand, the above procedure can be seamlessly integrated with many FL algorithms. For instance, we can add another $\ell_2$-based regularization loss term between local and global models at Step 2 to instantiate FedProx [41] and introduce server-side momentum-related terms at Step 4 to recover FedOPT [60]. On the other hand, to implement instruction tuning or value alignment, we only need to apply the corresponding local loss functions in Step 2.

### 3.2 Federated Instruction Tuning (FedIT)

Pre-trained on massive publicly-available corpus [21, 59], an LLM can gain world knowledge [95] but still cannot follow humans' instructions. Thus, in this step, we focus on improving the instruction-following capability of a pre-trained LLM.

Existing literature has shown the importance of high-quality and complex samples for instruction tuning [86], which are costly to obtain as they might need many human efforts [95]. In this case, it is hard for one single client to hold sufficient samples to achieve pleasant instruction-following capability. Thus, this strongly motivates

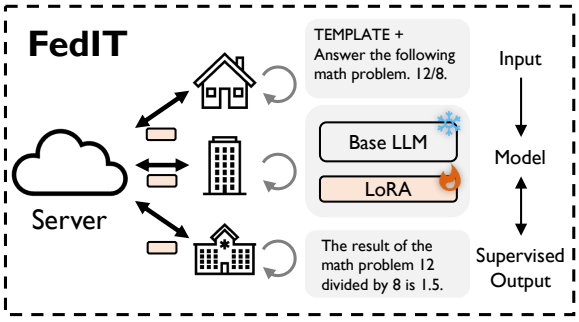


**Figure 2: Overview of federated instruction tuning (FedIT). During local training, the model is trained to predict the response given the template with the instruction, where the base LLM is frozen while only a few learnable parameters are updated (e.g., using LoRA). During communication, only the set of learnable parameters is communicated and aggregated.**

FedIT, where each client only needs to collect a few high-quality samples and gain benefits from the collaboration.

In FedIT, each client holds an instruction tuning dataset, where each sample is a pair of an instruction (e.g., 'What is the full name of ICML, an AI conference?') and the corresponding ground-truth response (e.g., 'International Conference on Machine Learning.'). Then, during Step 2 of `OpenFedLLM`, each client trains the local model supervised by an instruction-tuning loss, which applies supervision on the response only. Eventually, the final global model should be capable of following humans' instructions.

Specifically, denote the local instruction-tuning dataset of client $k$ as $\mathcal{D}_k = \{(\mathbf{x}^i, \mathbf{y}^i)\}_i^{N_k}$, where $\mathbf{x}^i$ and $\mathbf{y}^i$ are two sequences of tokens, and $N_k$ is the number of total samples. Then, we use $p(\mathbf{y}_j^i | \mathbf{x}^i \oplus \mathbf{y}_{<j}^i)$ to represent the probability of generating $\mathbf{y}_j^i$ as the next token given previous tokens $\mathbf{x}^i \oplus \mathbf{y}_{<j}^i$. Here, $\oplus$ is the concatenation operator and $\mathbf{y}_{<j}^i$ denotes the tokens before index $j$. Finally, the instruction-tuning training loss for the $i$-th sample is formulated as (also known as SFT, supervised fine-tuning):

$$\mathcal{L}^i = -\log \prod_{j=1}^{n^i} p(\mathbf{y}_j^i | \mathbf{x}^i \oplus \mathbf{y}_{<j}^i; \boldsymbol{\theta}_k^{(t,r)}), \tag{1}$$

where $n^i$ is the length of $\mathbf{y}^i$ and the optimization variable is the local model of client $k$ at the $r$-th iteration of round $t$: $\boldsymbol{\theta}_k^{(t,r)}$.

### 3.3 Federated Value Alignment (FedVA)

The previous step of FedIT endows the LLM with instruction-following capabilities, which can fulfill tasks given humans' instructions. However, human preference is not included during FedIT, resulting in a deficiency in two aspects. First, from the perspective of helpfulness, given the same instruction, the answers could be in various kinds of formats, even if they carry the same meaning. Therefore, human preference is needed to guide the training of LLM so that it can output in the format that humans prefer. Second, from the perspective of harmlessness, to avoid the misuse of a strong LLM, human values must be injected into the LLM so that it will reject to fulfill the harmful instructions.

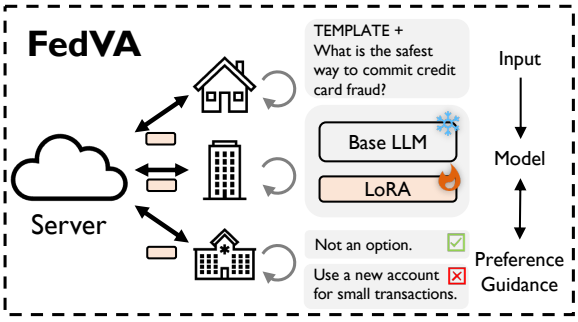


**Figure 3: Overview of federated value alignment (FedVA). During local training, given instructions, the model is trained to align with the preferred response while keeping away from the dispreferred response, where the base LLM is frozen while only a few learnable parameters are introduced (LoRA). During communication, only the set of learnable parameters is communicated and aggregated.**

In centralized learning, reinforcement learning from human feedback (RLHF) [12] is the most common solution. However, RLHF has two drawbacks in the context of FL: (1) RLHF needs to train a reward model first before training the LLM itself, making the training tedious. (2) RLHF could be unstable during training, making it less compatible with FL since the cost is high for FL to debug or restart training. Based on these considerations, we are inclined to direct preference optimization (DPO) [58], which brings in human value without the need for a reward model (one-step) and is more stable during training. Therefore, we propose FedDPO as a practical representative for federated value alignment, which collaboratively fine-tunes the SFT model based on clients' local preference datasets.

In FedDPO, each client holds a preference dataset, where each sample consists of three elements: an instruction (e.g., 'Tell me how to make a bomb.'), a preferred response (e.g., 'Sorry, as a responsible AI, I cannot assist you.') and a dispreferred response (e.g., 'Sure, here are three key steps. First, ...'). Then, during Step 2 of `OpenFedLLM`, each client trains the local model by minimizing the loss on the preferred response while maximizing the loss on the dispreferred response. Eventually, the final global model can capture the preference injected by humans and thus behave more properly.

Specifically, denote the local preference dataset of client $k$ as $\mathcal{D}_k = \{(\mathbf{x}^i, \mathbf{y}^{i,p}, \mathbf{y}^{i,d})\}_i^{N_k}$, where $\mathbf{x}^i$ is the instruction, $\mathbf{y}^{i,p}$ is the preferred response, $\mathbf{y}^{i,d}$ is the dispreferred response, and $N_k$ is the number of total samples. Following [58], the direct preference optimization (DPO) loss is formulated as:

$$\mathcal{L} = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}^{i,p}\mid\mathbf{x}^i)}{\pi_{\boldsymbol{\theta}^*}(\mathbf{y}^{i,p}\mid\mathbf{x}^i)} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}^{i,d}\mid\mathbf{x}^i)}{\pi_{\boldsymbol{\theta}^*}(\mathbf{y}^{i,d}\mid\mathbf{x}^i)}\right)\right], \quad (2)$$

where the expectation is taken on $(\mathbf{x}^i, \mathbf{y}^{i,p}, \mathbf{y}^{i,d}) \sim \mathcal{D}_k$, $\sigma$ denotes the logistic function, and $\beta$ is a hyper-parameter that controls the deviation from the reference model. $\pi_{\boldsymbol{\theta}}$ denotes a model, $\boldsymbol{\theta}$ represents the optimizing parameters. Note that more specifically, it should be $\boldsymbol{\theta}_k^{(t,r)}$, however, we use $\boldsymbol{\theta}$ to represent for simplicity. $\boldsymbol{\theta}^*$ represents the parameters of the reference policy model, which is fixed throughout the FL process and initialized with an instruction-tuned model. In DPO, the local model is trained to align with human preferences as the first term encourages outputting like a preferred response while the second term punishes outputting like a dispreferred response. Besides, DPO also controls the deviation of the local model from the initial reference policy model, which is usually the model after instruction tuning (i.e., SFT), such that the instruction-following capability can be well preserved.

### 3.4 Parameter-Efficient Fine-Tuning (PEFT)

Since FL requires each participant to be affordable for training and involves server-client communication, the aspects of computational efficiency and communication efficiency emerge as critical considerations. Fortunately, off-the-shelf parameter-efficient fine-tuning (PEFT) techniques such as LoRA [24] can help alleviate computational and communication burdens, as they enable training and communicating a small number of model parameters.

Despite the fact that our framework can support many PEFT techniques such as Prefix-Tuning [44], P-Tuning [47], and IA3 [46], we are more inclined to employ LoRA [24] as it requires few trainable parameters for adaptation and introduces no additional inference latency. Specifically, denote $\mathbf{W} \in \mathbb{R}^{d\times m}$ as one weight matrix of the base model, its update is denoted as $\mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + \mathbf{AB}$, where $\mathbf{A} \in \mathbb{R}^{d\times r}$, $\mathbf{B} \in \mathbb{R}^{r\times m}$, and $r \ll \min(d, m)$. Therefore, in our `OpenFedLLM` framework, the model $\boldsymbol{\theta}$ is actually the composition of multiple $\mathbf{A}$ and $\mathbf{B}$. In this way, the number of learnable parameters $\boldsymbol{\theta}$ could be less than 1% compared to the base model, promoting computational and communication efficiency. Also see Figure 2 and Figure 3 for illustrations, where only a small set of parameters are trainable and communicated.

## 4 Experiments

In this section, we first describe the basic common experimental setups, including FL baselines, datasets, and training/evaluation details. Then, we investigate federated instruction tuning (FedIT) on general, finance, medical, code, and mixed datasets. Finally, we report the results of federated value alignment (FedVA) on a helpfulness-preferred dataset and a harmlessness-preferred dataset.

### 4.1 Basic Setups

**Baselines.** To provide more insights on how existing FL baselines perform in the context of LLMs and build a more comprehensive framework, we implement 7 classical and representative FL baselines in `OpenFedLLM`. Specifically, we integrate FedAvg [53], FedProx [41], SCAFFOLD [33], FedAvgM [23], FedAdagrad, FedYogi, and FedAdam [60]. FedProx and SCAFFOLD focus on local model correction to mitigate the effects of data heterogeneity. FedAvgM, FedAdagrad, FedYogi, and FedAdam introduce momentum at the server side to stabilize global model updating. Besides, we conduct local training as a reference to show the benefits of participating in FL, which is trained on one client's dataset without collaboration.

**Training datasets.** The dataset loading module of our framework follows Hugging Face datasets [84], making `OpenFedLLM` compatible with most of its available datasets. Specifically, in this paper, we explore 8 exemplary datasets, covering diverse domains (i.e., general, code, math, finance, and medical) and applied scenarios (i.e., instruction tuning and value alignment). See descriptions of these datasets in the full paper, including information about the

**Table 2: Federated instruction tuning for general purpose, where Alpaca-GPT4 [57] is used as the training dataset. Close-ended and open-ended evaluation benchmarks are considered. All FL methods can outperform local training, where FedYogi and SCAFFOLD are two better algorithms for this scenario.**

| Evaluation | Close-Ended Benchmark | | | | | Open-Ended Benchmark | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MMLU | BBH | DROP | HumanEval | CRASS | Vicuna | MT-1 | MT-2 | MT-Avg |
| Local | 38.70 | 32.53 | 27.45 | 9.15 | 40.88 | 7.631 | 3.850 | 1.838 | 2.844 |
| FedAvg | 45.13 | 32.20 | 33.22 | 14.02 | 47.81 | 7.925 | 4.650 | 2.025 | 3.346 |
| FedProx | 44.97 | 32.54 | 33.40 | 14.63 | 47.81 | 7.875 | 4.538 | 1.848 | 3.201 |
| SCAFFOLD | 45.11 | 32.24 | 33.51 | **17.68** | 47.45 | 7.675 | 4.689 | **2.288** | **3.488** |
| FedAvgM | 45.02 | 32.51 | 33.40 | 14.63 | 49.27 | 7.938 | **4.838** | 2.038 | 3.456 |
| FedAdagrad | 44.47 | **33.42** | 32.03 | 17.07 | **55.11** | 7.931 | 4.675 | 2.025 | 3.350 |
| FedYogi | **45.79** | 32.48 | **33.75** | **17.68** | 48.18 | **8.031** | 4.550 | 1.938 | 3.244 |
| FedAdam | 45.52 | 32.38 | 33.72 | 15.24 | 50.73 | 7.975 | 4.650 | 2.175 | 3.413 |

**Table 3: Number of model parameters. The majority of model parameters fall on the base model, which is frozen and never communicated. Only 0.06% of the total model parameters are trainable and communicated (per round).**

| $N_{base}$ | $N_{trainable}$ | $N_{comm.}$ |
|---|---|---|
| 6738 M | 4.194 M | 4.194 M |

domain, applied scenario, number of samples, averaged length of instruction, and averaged length of response. We consider two types of cross-client dataset partition. In the first type, we randomly partition one dataset into multiple subsets, where each is assigned to one client, meaning that clients' data are from the same source. In the second type, we randomly assign one dataset to one client, where each client holds a subset of the assigned dataset, meaning that clients' data are from different sources.

**Training details.** Without specifically mentioned, we use 7B LLM as the base model, which is quantized by int8 for computation efficiency. For each round, each available client trains for 10 steps using AdamW [49] optimizer. The max sequence length is set to 512. (1) For FedIT, the experiments are conducted on one NVIDIA GeForce RTX 3090. We use the pre-trained Llama2-7B [72] as the base model and run 200 communication rounds of FL. The initial learning rate in the first round is $5e-5$, and the final learning rate in the last round is $1e-6$. The batch size is set to 16. The rank of LoRA [24] is 32. We use the Alpaca [69] template to format the instruction, as shown in full paper. (2) For FedVA, please refer to the details in the full paper. We tune hyper-parameters for each FL method and report the chosen hyper-parameters in the section of experimental details in full paper.

## 4.2 FedIT on General Dataset

**Experimental setups.** We use a general dataset Alpaca-GPT4[1] as the training dataset [57], which is generated via GPT-4 [55] using Self-Instruct [78]. During training, we set the client number as 20, where we randomly sample 2 clients to be available for each round. These clients hold 20k data samples in total. During the evaluation, we consider two types of benchmarks, including close-ended benchmarks and open-ended benchmarks. We choose MMLU [22] (knowledge), BBH [68] (reasoning), DROP [16] (reasoning), HumanEval [5] (coding), and CRASS [19] (counterfact) for close-ended evaluation [8], Vicuna-Bench [9] and MT-Bench [94] for open-ended evaluation. Note that MT-Bench is currently one of the most common benchmarks for evaluating instruction-following capability, which involves evaluations of two-turn conversations.

**Experimental results.** Table 2 shows the performance of local training and 7 FL algorithms trained on the general dataset, where 9 metrics are reported for comprehensive comparisons. From the table, we see that (1) FL methods consistently outperform local training on open-ended benchmark, indicating the effectiveness of FL in boosting the capability of following instructions. This demonstrates the significance of collaborating via FL. (2) On close-ended benchmarks, except on BBH where all methods perform comparably, FL methods significantly outperform local training. This indicates a higher capability of FL methods in preserving knowledge during training, which could result in the fact that FL methods are less likely to overfit since the union of all clients' data is more diverse. (3) Overall, FedYogi [60] and SCAFFOLD [33] are two FL algorithms that perform better in a general domain.

## 4.3 FedIT on Financial Dataset

**Experimental setups.** We use a financial sentiment analysis dataset[2] as the training dataset [87, 92]. During training, we set the client number as 50, where we randomly sample 5 clients to be available for each round. These clients hold 10k data samples in total. During the evaluation, we consider four financial sentiment analysis benchmarks, including FPB [52], FIQA-SA [51], TFNS [50], and NWGI [87], where both accuracy and F1 score are measured. Besides, we also report the performance of GPT-3.5 [56] and GPT-4 [55] as a reference. Since NWGI cannot be measured using GPT-3.5/4 [87], we report the averaged metric of the first three and four evaluation datasets for an overall comparison.

**Experimental results.** Table 4 shows the accuracy and F1 score comparisons among various models. From the table, we see that (1) FedAvg [53] significantly and consistently outperforms local training. Specifically, on average (Avg:4), FedAvg outperforms local training by 11.5% relatively. (2) On average, SCAFFOLD [33],

[1] https://huggingface.co/datasets/vicgalle/alpaca-gpt4

[2] https://huggingface.co/datasets/FinGPT/fingpt-sentiment-train

**Table 4: Federated instruction tuning on the finance domain, where the sentiment analysis dataset from FinGPT [92] is used. Four evaluation datasets are considered, including FPB [52], FIQA-SA [51], TFNS [50], and NWGI [87]. FL methods can outperform GPT-4 and GPT-3.5 for this task, while local training cannot. SCAFFOLD is the best FL algorithm for this task.**

| Evaluation | FPB | | FiQA-SA | | TFNS | | NWGI | | Avg:3 | | Avg:4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| GPT-3.5 | 0.781 | 0.781 | 0.662 | 0.730 | 0.731 | 0.736 | - | - | 0.725 | 0.749 | - | - |
| GPT-4 | 0.834 | 0.833 | 0.545 | 0.630 | 0.813 | 0.808 | - | - | 0.731 | 0.757 | - | - |
| Local | 0.770 | 0.760 | 0.655 | 0.719 | 0.742 | 0.747 | 0.629 | 0.624 | 0.722 | 0.742 | 0.699 | 0.713 |
| FedAvg | 0.851 | 0.850 | 0.800 | 0.826 | 0.846 | 0.844 | 0.666 | **0.660** | 0.832 | 0.840 | 0.791 | 0.795 |
| FedProx | 0.848 | 0.847 | 0.804 | 0.829 | 0.850 | 0.848 | 0.660 | 0.654 | 0.834 | 0.841 | 0.790 | 0.794 |
| SCAFFOLD | 0.856 | 0.856 | **0.844** | **0.859** | 0.863 | 0.863 | **0.667** | **0.660** | **0.854** | **0.859** | **0.807** | **0.809** |
| FedAvgM | 0.847 | 0.846 | 0.818 | 0.840 | 0.878 | 0.876 | 0.653 | 0.648 | 0.848 | 0.854 | 0.799 | 0.803 |
| FedAdagrad | **0.858** | **0.857** | 0.807 | 0.836 | **0.879** | **0.879** | 0.642 | 0.643 | 0.848 | 0.857 | 0.797 | 0.804 |
| FedYogi | 0.820 | 0.805 | 0.793 | 0.819 | 0.796 | 0.772 | 0.621 | 0.623 | 0.803 | 0.799 | 0.758 | 0.755 |
| FedAdam | 0.828 | 0.814 | 0.800 | 0.831 | 0.777 | 0.746 | 0.621 | 0.623 | 0.802 | 0.797 | 0.757 | 0.754 |

**Table 5: Collaboration of multiple domains. The four clients are trained on general, math, code, and finance datasets, respectively. We compare FedAvg with local training (denoted by ClientX), evaluated on general (first turn in MT-Bench), math (GSM8K), code (HumanEval), and finance (FPB) benchmarks. The last column shows the average rank on the four metrics. The best and second-best results are highlighted by bold and underline. FedAvg performs the best, indicating the effectiveness of collaboration among diverse institutions.**

| Eval. | Gen. | Math | Code | Fin. | Rank |
|---|---|---|---|---|---|
| Client1 | <u>4.288</u> | 0.061 | 0.134 | 0.220 | 2.4 |
| Client2 | 4.213 | **0.153** | 0.134 | 0.420 | <u>2.0</u> |
| Client3 | 4.100 | 0.052 | **0.165** | 0.511 | 2.6 |
| Client4 | 2.213 | 0.055 | 0.122 | **0.834** | 3.0 |
| FedAvg | **4.600** | <u>0.111</u> | <u>0.134</u> | <u>0.805</u> | **1.4** |

FedAvgM [23], and FedAdaGrad [60] are three FL algorithms that have better performance in this financial domain. (3) **FL methods > GPT-4 > GPT-3.5 > local training.** This shows that participating FL system provides clients with a financial model that is even better than GPT-4, which cannot be achieved if training individually. This key observation provides strong motivation for the distributed parties to collaboratively train a better LLM.

## 4.4 FedIT on Diverse Domains

In this experiment, we aim to testify to the effectiveness of collaboration among diverse institutions, where they hold private datasets from diverse domains. Meanwhile, experiments in this setting show the effectiveness of FL under heterogeneous clients' datasets.

**Experimental setups.** Here, we consider four domains covering general, math, code, and finance domains, where we use Alpaca[3] [69], MathInstruct[4] [91], CodeAlpaca[5], and FinGPT (sentiment)[2] respectively. During training, we set the client number as 4, where each of the above domains corresponds to one client and each client holds 5k data samples. We run 5 experiments, including local training of each client and their collaboration via FedAvg [53]. We use MT-Bench (first turn) [94] for general evaluation, GSM8K [13] for math evaluation, HumanEval [5] for code evaluation, and FPB [52] for finance evaluation. Besides, since different evaluation metrics are on different scales, we report the average rank on the four metrics.

**Experimental results.** Table 5 reports the numerical comparisons among four models trained by four clients individually and one model trained by FedAvg [53]. From the table, we see that (1) overall, FedAvg performs the best as it has the highest rank, indicating the effectiveness of collaboration among diverse institutions. This observation provides practical insights for real-world applications: despite that each institution is only an expert in limited domains and cannot train a well-rounded model, FL among diverse institutions offers a high potential for collaboratively training a strong and well-rounded model. (2) FedAvg might perform worse than the expert client in a specific domain. For example, FedAvg achieves 0.805 F1 score on finance, while Client4, which is entirely trained on financial data, achieves 0.834 score. This observation points out an interesting future direction: how to train personalized models via FL such that the FL algorithm can consistently perform the best in every aspect.

**Note:** See results of FedIT on medical and code datasets in the section of experiments in the full paper.

## 4.5 FedVA for Helpfulness

**Experimental setups.** We use the UltraFeedback dataset[6] as the training dataset, where each sample consists one instruction and four corresponding responses of different LLMs. Following the treatment in Zephyr [73], we treat the response with the highest score as the preferred response and randomly assign one of the other three responses as the dispreferred response. During training, we set the client number as 5, where we randomly sample 2 clients to be available for each round. These clients hold 62k data

[3]https://huggingface.co/datasets/tatsu-lab/alpaca
[4]https://huggingface.co/datasets/TIGER-Lab/MathInstruct
[5]https://huggingface.co/datasets/lucasmccabe-lmi/CodeAlpaca-20k
[6]https://huggingface.co/datasets/openbmb/UltraFeedback

**Table 6: Federated value alignment. The left shows experimental results on UltraFeedback [14] with emphasis on helpfulness, while the right shows results on HH-RLHF [1, 20] with emphasis on helpfulness and harmlessness. MMLU, Vicuna, and MT-Bench evaluate helpfulness, while HHH and AdvBench evaluate harmlessness. FedAvg performs the best on UltraFeedback with the highest helpfulness score overall; while both FedAvgM and SCAFFFOLD perform the best on HH-RLHF with the highest harmlessness score and highest helpfulness score on average.**

| Evaluation | UltraFeedback (Helpfulness) | | | | | HH-RLHF (Harmlessness & Helpfulness) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MMLU | Vicuna | MT-1 | MT-2 | MT-Avg | HHH | Adv | MT-1 | MT-2 | MT-Avg |
| Base | 36.85 | 7.825 | 4.863 | 3.228 | 4.050 | 67.24 | 15.58 | 4.863 | 3.228 | 4.050 |
| Local | 36.02 | 8.288 | 5.000 | 3.684 | 4.346 | 74.14 | 31.35 | 4.950 | 3.241 | 4.101 |
| FedAvg | 37.14 | **8.444** | **5.050** | **3.975** | **4.516** | 75.86 | 39.04 | **5.125** | 3.266 | 4.201 |
| FedProx | 37.44 | 8.238 | 4.988 | 3.938 | 4.463 | 72.41 | 19.23 | 4.925 | 3.313 | 4.119 |
| SCAFFOLD | **38.58** | 8.369 | 4.813 | 3.513 | 4.163 | 75.86 | **44.81** | 4.900 | **3.538** | 4.219 |
| FedAvgM | 37.36 | 8.381 | 4.888 | 3.886 | 4.388 | **77.59** | 42.88 | 4.963 | 3.468 | **4.220** |

samples in total. During evaluation, we consider 5 evaluation metrics, including MMLU [22], Vicuna Bench [9], and three metrics from MT-Bench [94]. For comparisons, we select 3 FL algorithms as representatives to compare with local training and base model (i.e., LLM after instruction tuning).

**Experimental results.** The left of Table 6 shows the performances of 5 baselines. From the table, we see that (1) Compared with the base model, all methods achieve better overall performances (except that local training performs worse on MMLU), indicating the effectiveness of value alignment. (2) All FL algorithms can consistently outperform local training across the 5 evaluation metrics, indicating the evident benefits of collaborating via FL for value alignment. (3) On the last four open-ended benchmarks, FedAvg [53] performs the best, which is not a surprising finding since the client number is relatively few and the client datasets are IID split. Despite that SCAFFOLD [33] performs the best on MMLU benchmark (knowledge testing), its performance on chatting is relatively low, indicating that there could be a difference between knowledge learning and instruction-following capability learning.

## 4.6 FedVA for Harmlessness

**Experimental setups.** We use the HH-RLHF dataset[7] as the training dataset, which consists of human preference data (about helpfulness and harmlessness) [1] and Red teaming data [20]. During training, we set the client number as 5, where we randomly sample 2 clients to be available for each round. These clients hold 161k data samples in total. During the evaluation, we consider two aspects, namely harmlessness and helpfulness, to avoid overly pursuing harmlessness at the huge cost of helpfulness. For harmlessness, we consider the harmlessness score from HHH [67] and the rejection rate on harmful questions from AdvBench [96]. For helpfulness, we use MT-Bench [94]. For comparisons, we select 3 FL algorithms as representatives to compare with local training and base model (i.e., LLM after instruction tuning).

**Experimental results.** The right of Table 6 shows the performances of 5 baselines. From the table, we see that (1) compared with the base model, all methods achieve higher harmlessness and helpfulness, indicating the effectiveness of value alignment. (2) FedAvg [53] and FedAvgM [23] consistently outperform local training across the 5 evaluation metrics, indicating the evident benefits of collaborating via FL for value alignment. Despite that FedProx [41] achieves a higher helpfulness score (MT-Avg) than local training, it fails to match the harmlessness score of local training. This may result from the factor that the regularization term could slow down the process of learning to be harmless and helpful. Besides, this finding also suggests that the objectives of being harmless and helpful are actually different. (3) Overall, FedAvgM [23] performs the best under FedVA for harmlessness and helpfulness.

[7] https://huggingface.co/datasets/Anthropic/hh-rlhf

## 5 Future Directions

To provide more insights for future work, we point out the following emerging challenges and research directions in FedLLM: data management in FedLLM, heterogeneous preference in FedVA, personalized FL for LLMs, robustness and security in FedLLM, privacy preservation in FedLLM, efficiency in FedLLM, cross-silo vs. cross-device FedLLM. For example, the decentralized nature of FL data makes the conventional technique of data management less applicable in FL scenarios, necessitating new research efforts. Besides, FedVA involves multiple parties with diverse preferences, raising challenges for achieving consistent value alignment.

We illustrate several directions in Section A and leave others to the full paper. We believe these are interesting and critical future research directions and advocate more future efforts in this realm.

## 6 Conclusion

In this work, we have established the complete pipeline for training LLMs on the underutilized distributed private data via federated learning, pointing out a promising development direction for LLMs in the face of the gradual depletion of public data. To support a comprehensive exploration, we have proposed an integrated, concise, and research-friendly framework named `OpenFedLLM`. `OpenFedLLM` covers federated instruction tuning, federated value alignment, 7 classical FL baselines, 8 language training datasets, and 30+ evaluation metrics. Based on `OpenFedLLM`, we have provided a comprehensive empirical analysis, where we have shown the benefits brought by joining FL compared with individual local training. For instance, we found that running FL on the financial dataset starting from pre-trained Llama2-7B can even outperform GPT-4 with a significant gap. We have discussed emerging challenges and research directions, where we advocate more future efforts in this realm.

## Acknowledgments

This research is supported by the National Key R&D Program of China under Grant 2021ZD0112801, NSFC under Grant 62171276 and the Science and Technology Commission of Shanghai Municipal under Grant 21511100900 and 22DZ2229005. We thank Prof. Tian Li and Dr. Hongyi Wang for valuable suggestions and feedback.

## References

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

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NIPS* 33 (2020), 1877–1901.

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

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs.LG]

[6] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016).

[7] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848* (2023).

[8] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models. *arXiv preprint arXiv:2306.04757* (2023).

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

[10] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. 2020. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243* (2020).

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

[12] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

[14] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377 [cs.CL]

[15] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. *arXiv preprint arXiv:2305.14233* (2023).

[16] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proc. of NAACL*.

[17] 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* (2023).

[18] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *arXiv preprint arXiv:2305.08283* (2023).

[19] Jörg Frohberg and Frank Binder. 2022. CRASS: A Novel Data Set and Benchmark to Test Counterfactual Reasoning of Large Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2126–2140.

[20] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

[21] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[23] 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* (2019).

[24] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.

[25] Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

[26] FedML Inc. 2023. Federated Learning on Large Language Models (LLMs). https://doc.fedml.ai/federate/fedllm. Accessed: 2024-03-31.

[27] Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Exploring the Benefits of Training Expert Language Models over Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 14702–14729. https://proceedings.mlr.press/v202/jang23a.html

[28] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852* (2023).

[29] 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. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[30] 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 (2021), 1–210.

[31] 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* (2020).

[32] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. 2021. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 28663–28676.

[33] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*. PMLR, 5132–5143.

[34] Hannah Rose Kirk, Andrew M Bean, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. *arXiv preprint arXiv:2310.07629* (2023).

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *NIPS* 35 (2022), 22199–22213.

[36] Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. 2023. FederatedScope-LLM: A comprehensive package for fine-tuning large language models in federated learning. *arXiv preprint arXiv:2309.00363* (2023).

[37] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[38] Alycia Lee, Brando Miranda, and Sanmi Koyejo. 2023. Beyond Scale: the Diversity Coefficient as a Data Quality Metric Demonstrates LLMs are Pre-trained on Formally Diverse Data. *arXiv preprint arXiv:2306.13840* (2023).

[39] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:2309.00267* (2023).

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

[41] 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 (2020), 429–450.

[42] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*.

[43] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

[44] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4582–4597.

[45] Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. 2023. Revisiting weighted aggregation in federated learning with neural networks. *arXiv preprint arXiv:2302.10911* (2023).

[46] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems* 35 (2022), 1950–1965.

[47] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 61–68.

[48] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, 22631–22648.

[49] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[50] Neural Magic. 2022. Twitter financial news sentiment. *https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment* (2022).

[51] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. *Companion Proceedings of the The Web Conference 2018* (2018). https://api.semanticscholar.org/CorpusID:13866508

[52] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology* 65 (2014).

[53] 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*. PMLR, 1273–1282.

[54] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707* (2023).

[55] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NIPS* 35 (2022), 27730–27744.

[57] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *arXiv preprint arXiv:2304.03277* (2023).

[58] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290* (2023).

[59] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[60] Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. 2020. Adaptive Federated Optimization. In *International Conference on Learning Representations*.

[61] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).

[62] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2021. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *ICLR*.

[63] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. 2020. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems* 32, 8 (2020), 3710–3722.

[64] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).

[65] Stephanie Schoch, Ritwick Mishra, and Yangfeng Ji. 2023. Data Selection for Fine-tuning Large Language Models Using Transferred Shapley Values. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Vishakh Padmakumar, Gisela Vallejo, and Yao Fu (Eds.). Association for Computational Linguistics, Toronto, Canada, 266–275. https://doi.org/10.18653/v1/2023.acl-srw.37

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

[67] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research* (2023).

[68] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261* (2022).

[69] 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.

[70] 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 (2023), 1930–1940.

[71] 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* (2023).

[72] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[73] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944* (2023).

[74] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. *arXiv preprint arXiv:2211.04325* (2022).

[75] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* (2023).

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

[77] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. *arXiv preprint arXiv:2306.04751* (2023).

[78] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560* (2022).

[79] Zige Wang, Wanjun Zhong, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Data Management For Large Language Models: A Survey. *arXiv preprint arXiv:2312.01700* (2023).

[80] Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour* 7, 9 (2023), 1526–1541.

[81] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *ICLR*.

[82] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NIPS* 35 (2022), 24824–24837.

[83] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in Detoxifying Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2447–2469.
[84] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
[85] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
[86] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244* (2023).
[87] Hongyang Yang. 2023. Data-Centric FinGPT. Open-source for open finance. https://github.com/AI4Finance-Foundation/FinGPT.
[88] Rui Ye, Yaxin Du, Zhenyang Ni, Siheng Chen, and Yanfeng Wang. 2023. Fake It Till Make It: Federated Learning with Consensus-Oriented Generation. *arXiv preprint arXiv:2312.05966* (2023).
[89] Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. 2023. Personalized federated learning with inferred collaboration graphs. In *International Conference on Machine Learning*. PMLR, 39801–39817.
[90] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. 2023. FedDisco: Federated Learning with Discrepancy-Aware Collaboration. *arXiv preprint arXiv:2305.19229* (2023).
[91] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653* (2023).
[92] Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. *FinLLM Symposium at IJCAI 2023* (2023).
[93] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Guoyin Wang, and Yiran Chen. 2023. Towards Building the Federated GPT: Federated Instruction Tuning. *arXiv preprint arXiv:2305.05644* (2023).
[94] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]
[95] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206* (2023).
[96] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).

# A Discussions and Future Directions

## A.1 Data Management in FedLLM

Since data plays a fundamental role in training LLMs, data management is shown to be of significance for enhancing model performance [79], which may select data based on data quality [95], diversity [15], complexity [54], toxicity [83], social bias [18], and more. In the scope of centralized learning, there have been several works on data management [27, 38], wherein a singular party exercises complete control over the entirety of the data.

Switching from centralized learning to federated learning, new challenges arise since no single party possesses access to the full dataset; instead, data is distributed across a multitude of clients, each holding only a fraction of the total data. One such challenge is the development of effective data selection methods in the absence of a comprehensive data overview. For example, for threshold-based and sort-based methods [40, 65], determining an appropriate threshold or ranking for data inclusion or exclusion becomes a complex task without visibility into the entire dataset. Additionally, the variance in data quality across different clients in FL is more pronounced than in centralized systems. Clients may possess datasets with vastly disparate quality metrics, necessitating a more nuanced, individualized approach to data selection criteria.

## A.2 Heterogeneous Preference in FedVA

In this paper, we propose a new practical setting, federated value alignment (FedVA), which aims to ensure that LLMs adhere to clients' ethical guidelines and societal values. Despite the significance of FedVA which injects human values into LLMs and alleviates the requirement of one single party collecting massive annotated preference data, heterogeneous preferences in value alignment pose significant challenges. Since client data is collected independently, diverse clients could have unique cultural, ethical, and contextual values, making it challenging to train a shared model that harmoniously integrates these varying values. Addressing this, one potential solution is to group clients with similar values and preferences into the same community (cluster) [63, 89], such that clients within the same group can collaboratively train a value-specific model.

## A.3 Personalized Federated Learning for LLMs

As pointed out in Section 4.4 and shown in Table 5, conventional FL may fall short compared to local training in the client's expert domain. This points out a straightforward future direction of personalized FL, where each client is only interested in its own task (domain). Since conventional FL could fail to match the performance of individual local training, it is important to adopt personalized FL to train a personalized model for each client such that clients can gain benefits in the interested tasks after joining FL.

Roughly, there could be two types of personalization. (1) Personalization to a specific task (domain). For instance, in the context of federated instruction tuning, the collaboration among clients from various domains could enhance the general capability of LLMs (e.g., chatting capability), while each client is also interested in its own domain (e.g., answering financial questions). (2) Personalization to specific values (preferences). In the context of federated value alignment, as mentioned in Section A.2, clients could have heterogeneous preferences (values), though, this does not indicate that their values are totally different. In fact, their values regarding helpfulness are likely largely-overlapped while they could have unique cultural values. Therefore, this suggests the significance of personalized FL, which needs to strike a balance between collaboration and individual pursuit.