CoFL: Consensus Driven Human-AI Collaborative Federated Learning

Zeyuan Cai[1], **Yao Zhang**[1*], **Zhiwen Yu**[2,1*], **Jiaqi Liu**[1], **Yuchang Sun**[3], **Chenhao Ma**[1] and **Yilin Zhao**[1]
[1]Northwestern Polytechnical University
[2]Harbin Engineering University
[3]Independent Researcher
{zy_cai, 2025263083, 13176361085}@mail.nwpu.edu.cn, {yaozh.g, zhiwenyu, jqliu}@nwpu.edu.cn

Abstract

Federated learning (FL) has emerged as a distributed machine learning paradigm due to its privacy-preserving advantages. Most FL studies assume offline labeled datasets are available at clients. In practice, however, client data often arrive without labels in a streaming manner, making label acquisition a crucial problem. Existing solutions typically rely on human experts or artificial intelligence (AI) models for annotation. Nevertheless, due to the subjectivity among human experts and the inherent limitations of AI models, the labels they provide are often unreliable. In this work, we propose CoFL, a novel consensus driven human-AI collaborative FL method, which utilizes the complementarity between humans and AI models to produce more reliable labels. Moreover, CoFL leverages cross-client consensus among human experts to further enhance the collaboration process. Experiments on three datasets demonstrate that CoFL consistently outperforms baseline methods under various settings. The code is available at https://github.com/huaiguang233/CoFL.

1 Introduction

Federated Learning (FL) enables multiple clients to collaboratively train a global model without sharing raw data [Wang *et al.*, 2025a; Wen *et al.*, 2023]. Due to its privacy preserving property, FL has attracted widespread attention across various domains, including healthcare [Wu *et al.*, 2022], finance [Zhao *et al.*, 2023], and recommender systems [Wang *et al.*, 2025b]. Despite the success of FL, most methods assume a labeled dataset is available at each client [Tang and Yu, 2025; Shao *et al.*, 2024]. In practice, data often arrive at clients in an unlabeled streaming fashion [Sun *et al.*, 2025; Yu *et al.*, 2021]. For example, hospitals constantly accumulate raw diagnostic data that often lack labels. The absence of labels prevents local training and hinders the deployment of FL.

Recent efforts that focus on this issue roughly fall into Federated Semi-Supervised Learning (FSSL) and Federated Active Learning (FAL). Specifically, FSSL integrates semi-supervised learning to generate pseudo-labels [Diao *et al.*, 2022; Lee *et al.*, 2024], while FAL queries local human expert to annotate informative subsets of local unlabeled data at each client [Chen *et al.*, 2024; Sun *et al.*, 2025].

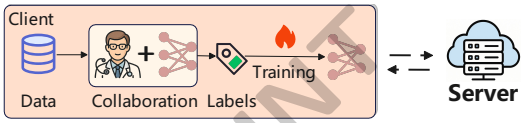


Figure 1: Workflow of proposed CoFL.

Despite their promise, the practical adoption of the above solutions remains challenging. FSSL often struggles with the inaccurate pseudo-labels generated by global or local *models* [Zhao *et al.*, 2025a]. Meanwhile, most FAL methods assume *human experts* are oracle annotators and overlook annotation inconsistencies caused by the subjectivity of experts. Such subjectivity leads to differences in experts' *capability*, i.e., their accuracy in annotating a specific class. This limitation is particularly severe in FL where experts are distributed across different clients and may have different educational backgrounds and preferences. Overall, labels from AI models and human experts are often unreliable.

Recent studies on *human-AI complementarity* [Vaccaro *et al.*, 2024; Steyvers *et al.*, 2022; Zöller *et al.*, 2025] demonstrate that human-AI hybrid systems can outperform either humans or AI models alone. These findings motivate a potential solution: leveraging human-AI collaboration to generate more reliable labels. Specifically, we consider a new FL method where each client pairs a human expert with the global or local model to collaboratively label incoming data, as illustrated in Fig. 1.

Nevertheless, implementing the aforementioned potential solution still presents several challenges. (1) *How to achieve effective human-AI collaboration?* Most existing human-AI collaboration approaches assume that the AI model is well trained [Kerrigan *et al.*, 2021; Zhang *et al.*, 2024]. In contrast, in the early stages of FL, limited available data leads to insufficiently trained models. (2) *How to identify samples for querying human experts?* Constrained by limited time and budget, it is more practical to query only a subset of unlabeled data for human annotations. Although sample querying strategies have been widely studied in the field of FAL, existing FAL methods [Kim *et al.*, 2023;

*Corresponding authors.

Sun *et al.*, 2025] cannot be directly applied because they ignore expert subjectivity. To address the above challenges, our idea is to leverage consensus among human experts to guide human-AI collaboration and sample querying. Formally, we define consensus as the majority-vote label among all human experts for the same sample. Existing studies [Raykar *et al.*, 2010] have shown that consensus among experts is more robust and closer to the ground truth label, which can be used to stabilize human-AI collaboration and improve sample querying by calibrating experts' subjectivity. To observe consensus across experts, existing methods assume that each sample is annotated by multiple experts [Showalter *et al.*, 2024; Kelly *et al.*, 2025]. This assumption, however, does not hold in privacy-preserving FL settings, where each client only has access to annotations from its local expert. Consequently, *how to estimate cross-client human expert consensus under privacy constraints* remains a critical challenge.

In this paper, we propose **CoFL** (Consensus-driven Human-AI Collaborative Federated Learning), a novel FL method that establishes a collaboration mechanism between human experts and AI models for reliable label generation under streaming unlabeled data. To tackle the aforementioned challenges, the key idea of CoFL is to learn knowledge from the historical annotations of human experts and aggregate such knowledge at the server to form cross-client consensus, which guides both human-AI collaboration and sample querying. Specifically, CoFL employs an Evidential Deep Learning (EDL) model [Sensoy *et al.*, 2018] on each client to learn the labeling behavior of experts influenced by subjectivity and aggregates the EDL models at the server into a consensus estimator. Moreover, CoFL calibrates the EDL outputs with the Expectation-Maximization (EM) algorithm, enabling the consensus to guide collaboration and sample querying more effectively.

Our main contributions can be summarized as follows:

- We propose a novel human-AI collaborative FL method CoFL, which utilizes the collaboration between human experts and models to generate reliable labels.
- We introduce a privacy-preserving consensus estimation method that constructs cross-client human expert consensus without sharing raw data or annotations. Meanwhile, we further introduce a calibration mechanism based on the EM algorithm, which enables the estimated consensus to better improve the effectiveness of human-AI collaboration.
- Experimental results demonstrate that, compared with five baseline methods, CoFL generates more reliable labels and significantly improves the performance of global model on three datasets.

2 Related Work

2.1 Federated Learning with Unlabeled Data

Existing FL methods with unlabeled data can be broadly categorized into two classes: FSSL and FAL. FSSL leverages global or local models to generate pseudo-labels for self-training [Jiang *et al.*, 2024; Diao *et al.*, 2022]. For example, SAGE [Liu *et al.*, 2025] leverages the confidence discrepancy between the local model and the global model to assess the reliability of pseudo-labels and accordingly correct them. In contrast, FAL incorporates active learning into FL and queries human experts for the most informative samples under a limited labeling budget [Cao *et al.*, 2023; Chen *et al.*, 2024]. For example, LoGo [Kim *et al.*, 2023] proposes a two-stage sampling strategy, in which local models are first used to cluster local samples, and then the global model is employed to identify the samples with highest uncertainty within each cluster. In this way, LoGo achieves robust and efficient sample querying.

The work most relevant to ours is [Sun *et al.*, 2025], which formulates the sample querying problem as a partially observable Markov decision process and solves it with multi-agent reinforcement learning. Compared to [Sun *et al.*, 2025], we consider a more realistic scenario in which the human experts are subjective and may provide inconsistent annotations.

2.2 Human-AI Collaboration

Although AI has made great breakthroughs in recent years, it is widely acknowledged that AI models still exhibit inherent weaknesses [Geirhos *et al.*, 2020]. In contrast, humans retain unique advantages in perception and reasoning [Yax *et al.*, 2024]. Thus, human-AI collaboration has emerged as a promising paradigm for addressing complex problems. Existing human-AI collaboration methods can be broadly categorized into two paradigms [Zhang *et al.*, 2024; Zhao *et al.*, 2025b]: Learning to Complement (L2C) and Learning to Defer (L2D). L2C focuses on fusing human and AI outputs to produce superior results. For example, [Steyvers *et al.*, 2022] proposed a Bayesian fusion framework that integrates human and AI predictions. [Kerrigan *et al.*, 2021] incorporates confusion matrices and calibration mechanisms for both humans and models, enabling effective human-AI collaboration. L2D aims to improve human-AI collaboration performance by assigning tasks to either the AI model or human expert. For instance, [Madras *et al.*, 2018] models the AI and the human expert as a two-stage collaborative decision framework, in which the AI not only needs to learn how to predict the class of the sample, but also needs to learn whether to defer the sample to the human expert. Furthermore, [Zhang *et al.*, 2024] extends the L2D framework to multi-expert settings by jointly learning when to defer and how many experts to involve.

3 System Overview

3.1 System Setting

We consider an FL system consisting of a central server and K clients that collaboratively train a global model $\theta_{\mathrm{g}} \in \mathbb{R}^d$ for a C-class classification task. We model the entire FL process as T discrete time steps. At each time step $t \in \{1, 2, \ldots, T\}$, each client $k \in [K]$ receives a newly arrived unlabeled dataset $\mathcal{D}_k^t$, where each sample $x \in \mathcal{D}_k^t$ is drawn from a distribution P_k. We assume $\bigcup_{k\in[K]} P_k = P_{\mathrm{glob}}$, while $P_k \neq P_{\mathrm{glob}}$ for all $k \in [K]$, reflecting the non-independent and identically distributed (Non-IID) nature of data across clients[Li *et al.*, 2022]. After generating labels for the newly arrived samples, clients participate in B FL communication

rounds at the current time step to collaboratively train the global model.

Human Experts and Sample Querying. We assume that each client is associated with a human expert, denoted by h_k. At time step t, the client k selects a subset of samples $\mathcal{Q}_k^t \subseteq \mathcal{D}_k^t$ to query its associated expert h_k for annotations and the query strategy is denoted by φ. Due to constraints on time and budget, the query budget is limited by $|\mathcal{Q}_k^t| \leq n_k$. For any sample $x \in \mathcal{Q}_k^t$, the expert h_k provides an annotation $h_k(x) \in \{1, \ldots, C\}$. The human annotation set is denoted as $\mathcal{U}_k^t = \{(x, h_k(x)) \mid x \in \mathcal{Q}_k^t\}$. φ will be detailed in Sec. 4.2.

Consensus. For a given sample x, we define the *consensus* as the majority class determined by aggregating expert-provided labels across all clients. Specifically, suppose that x could be annotated by the expert associated with each client. We treat each annotation as a vote and define the vote count histogram as $\mathbf{H}(x) = \sum_{k=1}^{K} \text{one-hot}(h_k(x)) \in \mathbb{N}^C$, where $\text{one-hot}(\cdot)$ denotes the one-hot encoding. The consensus is then given by $y^\star(x) := \arg\max_{c \in [C]} \mathbf{H}_c(x)$. To probabilistically characterize expert consensus, we model the latent consensus distribution with a Dirichlet distribution and the aggregated expert votes with a Multinomial distribution conditioned on it. Formally, we state the following assumption:

Assumption 1. *For any sample x, there exists a latent human consensus distribution $\boldsymbol{\pi}(x) \in \Delta^{C-1}$ such that*

$$\begin{aligned} \boldsymbol{\pi}(x) &\sim \text{Dirichlet}(\boldsymbol{\alpha}^\star(x)), \\ \mathbf{H}(x) &\sim \text{Multinomial}(K, \boldsymbol{\pi}(x)), \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha}^\star(x) \in \mathbb{R}_+^C$ denotes the unknown Dirichlet concentration parameter.

Since each client's local data cannot be exposed to experts from other clients, directly observing the consensus is not feasible. We introduce a *consensus estimator* $g(x; w)$ that estimates the consensus for arbitrary samples x under privacy constraints and will be detailed in Sec. 4.1.

Consensus-enhanced Human-AI Collaboration. At time step t, each client performs label generation for its newly arrived data $\mathcal{D}_k^t$ by leveraging three sources: (1) the global model performs inference[1] on all samples $x \in \mathcal{D}_k^t$, yielding a set of model predictions denoted as $\mathcal{I}_k^t = \{(x, f(x; \theta_{\text{g}}^{t-1})) \mid x \in \mathcal{D}_k^t\}$, where $f(x; \theta_{\text{g}}^{t-1}) \in \Delta^{C-1}$ represents the class probability vector over the C classes for sample x, θ_{g}^{t-1} denotes the global model obtained after the previous time step; (2) the consensus estimator produces a set of estimated consensus, denoted as $\mathcal{O}_k^t = \{(x, g(x; w)) \mid x \in \mathcal{D}_k^t\}$; (3) human expert annotations on the queried subset $\mathcal{Q}_k^t$ yield a labeled set $\mathcal{U}_k^t = \{(x, h_k(x)) \mid x \in \mathcal{Q}_k^t\}$.

The fusion of these three sources forms a consensus-enhanced human-AI collaboration process, denoted by $\Phi(\mathcal{I}_k^t, \mathcal{O}_k^t, \mathcal{U}_k^t)$, which assigns each sample $x \in \mathcal{D}_k^t$ a label y' along with a confidence s. Samples whose confidence exceeds a predefined threshold τ are retained to form $\mathcal{M}_k^t$, which is then added to the historical labeled set $\mathcal{L}_k^t$ for subsequent local training, i.e., $\mathcal{L}_k^t = \mathcal{L}_k^{t-1} \cup \mathcal{M}_k^t$. $\Phi(\mathcal{I}_k^t, \mathcal{O}_k^t, \mathcal{U}_k^t)$ will be detailed in Sec. 4.3.

[1]Alternatively, the local model can also be used for inference. Using the global model is a more common choice [Sun *et al.*, 2025].

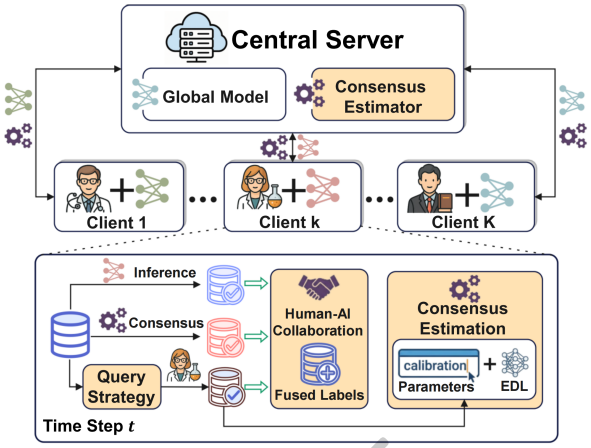


Figure 2: Framework of CoFL. At each time step, after receiving arrived unlabeled data, CoFL selects a subset for human annotation, while the global model and consensus estimator produce predictions and estimated consensus for all samples. These three sources are fused to generate reliable labels for local training and aggregation. Meanwhile, the newly collected human annotations are used to update the consensus estimator for the next time step.

3.2 Problem Formulation

We adopt FedAvg [McMahan *et al.*, 2017] for local updates and global aggregation. At the beginning, the server initializes a global model θ_{g}^0 and distributes it to all clients, each of which has a small labeled set $\mathcal{L}_k^0$ for initialization. At time step t, the previous global model θ_{g}^{t-1} initializes the current training process, which optimizes the following objective:

$$\min_{\theta_{\text{g}}^t} F(\theta_{\text{g}}^t) \triangleq \sum_{k=1}^{K} \frac{|\mathcal{L}_k^t|}{|\mathcal{L}^t|} F_k(\theta_{\text{g}}^t; \mathcal{L}_k^t), \tag{2}$$

where $\mathcal{L}^t = \bigcup_{k=1}^{K} \mathcal{L}_k^t$ denotes the union of labeled datasets across all clients at time step t.

The local loss function of client k is defined as

$$F_k(\theta_{\text{g}}^t; \mathcal{L}_k^t) \triangleq \frac{1}{|\mathcal{L}_k^t|} \sum_{(x, y') \in \mathcal{L}_k^t} \ell(\theta_{\text{g}}^t; x, y'),$$

$$\begin{aligned} \text{where} \quad \mathcal{L}_k^t &= \mathcal{L}_k^{t-1} \cup \mathcal{M}_k^t, \\ \mathcal{M}_k^t &= \left\{(x, y') \,\middle|\, (x, y', s) \in \Phi(\mathcal{I}_k^t, \mathcal{O}_k^t, \mathcal{U}_k^t),\ s > \tau\right\}, \\ \mathcal{I}_k^t &= \left\{(x, f(x; \theta_{\text{g}}^{t-1})) \mid x \in \mathcal{D}_k^t\right\}, \\ \mathcal{O}_k^t &= \left\{(x, g(x; w)) \mid x \in \mathcal{D}_k^t\right\}, \\ \mathcal{U}_k^t &= \left\{(x, h_k(x)) \mid x \in \mathcal{Q}_k^t, \mathcal{Q}_k^t = \varphi(\mathcal{D}_k^t)\right\}, \end{aligned} \tag{3}$$

where $\ell(\cdot)$ is a task-specific loss function.

Accordingly, our goal is to design the fusion strategy Φ, the consensus estimator $g(\cdot; w)$, and the query strategy φ such that each client can construct a reliable labeled set $\mathcal{L}_k^t$.

4 Methodology

In this section, we present CoFL in detail, which consists of three main modules: consensus estimation, consensus enhanced human-AI collaboration, and a capability-aware sample querying strategy. The framework is shown in Fig. 2.

4.1 Consensus Estimation

To estimate human expert consensus at each client, we propose a privacy-preserving consensus estimation method. The key idea is to learn a global surrogate model of human labeling behavior through federated aggregation, such that the aggregated surrogate captures an implicit consensus among distributed experts [Luo *et al.*, 2023]. We first construct a *surrogate model* for human experts that can generate annotations without exposing raw data. Prior studies have shown that human decision-making is inherently uncertain [Peterson *et al.*, 2019], while conventional neural networks typically produce point estimates and do not explicitly characterize prediction uncertainty [Gal and Ghahramani, 2016]. We model expert labeling behavior as a stochastic process that samples from a Dirichlet distribution [Yang *et al.*, 2025; Muttenthaler *et al.*, 2025]. For a given sample x, the labeling behavior of the h_k satisfies:

$$\begin{aligned} h_k(x) &\sim \text{Categorical}(\boldsymbol{\pi}_k(x))\,, \\ \boldsymbol{\pi}_k(x) &\sim \text{Dirichlet}(\boldsymbol{\beta}_k(x))\,. \end{aligned} \tag{4}$$

Note that $\boldsymbol{\beta}_k(x)$ and $\boldsymbol{\alpha}^\star(x)$ in Assumption 1 have different meanings: $\boldsymbol{\beta}_k(x)$ represents the subjectivity and uncertainty of an individual expert h_k, whereas $\boldsymbol{\alpha}^\star(x)$ characterizes the subjectivity and uncertainty of consensus among all experts.

To learn the Dirichlet parameters from experts, we further introduce an EDL model at client as surrogate for h_k. The EDL model parameterizes a Dirichlet-Categorical distribution by predicting Dirichlet parameters [Sensoy *et al.*, 2018]. For any sample x, the EDL outputs the Dirichlet parameters, denoted as $\hat{\boldsymbol{\beta}}_k(x) = [\hat{\beta}_{k,1}(x), \ldots, \hat{\beta}_{k,C}(x)] \in \mathbb{R}_+^C$, where $\hat{\beta}_{k,c}(x) = e_c(x) + 1$. The $e_c(x)$ quantifies the amount of evidence supporting class c, while the additive constant 1 represents a uniform Dirichlet prior, ensuring a valid distribution even when no evidence is observed [Sensoy *et al.*, 2018].

At time step t, for client k, the historical annotations provided by its human expert are denoted by $\mathcal{U}_k^{\leq t} = \bigcup_{i=1}^{t} \mathcal{U}_k^i$. Within the time step, the EDL model is trained in a federated manner together with the task model for B communication rounds. Specifically, at each communication round, client k updates its local EDL model $g(\cdot; w_k^t)$ using $\mathcal{U}_k^{\leq t}$, where w_k^t denotes the parameters of the local EDL model. The server then aggregates the local EDL models from participating clients to obtain a global EDL model $g(\cdot; w_g^t)$, where w_g^t denotes the parameters of the global EDL. The detailed training procedure is provided in Appendix A. After B communication rounds, the global EDL model serves as consensus estimator at next time step. For any sample x, each client can locally infers the Dirichlet parameters of consensus, denoted as $\hat{\boldsymbol{\beta}}_g(x)$.

4.2 Capability-Aware Sample Query Strategy

In this section, we introduce a capability-aware sample querying strategy. The key idea is to utilize the discrepancy between local and global EDL models to assess whether each sample aligns with the capability of the current human expert. While the global EDL model from the last time step $g(\cdot; w_g^{t-1})$ captures the consensus of human experts across all clients, each local EDL model $g(\cdot; w_k^{t-1})$ reflects the labeling behaviors of h_k. The discrepancy between the local and global EDL predictions therefore provides an effective signal for assessing whether a sample is aligned with the expert's capability, i.e., whether it belongs to a category with high expert annotation accuracy.

For a sample $x \in \mathcal{D}_k^t$, the client first infers the outputs from both the local and global EDL models: $\hat{\boldsymbol{\beta}}_k(x) = g(x; w_k^{t-1})$, $\hat{\boldsymbol{\beta}}_g(x) = g(x; w_g^{t-1})$. We then compute the expected class probability distributions induced by the corresponding Dirichlet parameters, denoted as $\mathbf{p}_k(x) = \frac{\hat{\boldsymbol{\beta}}_k(x)}{\sum_{j=1}^{C} \hat{\beta}_{k,j}(x)}$, $\mathbf{p}_g(x) = \frac{\hat{\boldsymbol{\beta}}_g(x)}{\sum_{j=1}^{C} \hat{\beta}_{g,j}(x)}$. We quantify the capability mismatch between the local expert and the global consensus using a distributional discrepancy measure $d_k(x) = KL(\mathbf{p}_g(x), \mathbf{p}_k(x))$, where $KL(\cdot,\cdot)$ is Kullback-Leibler (KL) divergence[2], which quantifies the discrepancy between two probability distributions. A larger discrepancy indicates that the local expert's belief on x deviates significantly from the global consensus, suggesting a higher risk of unreliable annotation.

Based on the discrepancy $d_k(x)$, we select the n_k samples from $\mathcal{D}_k^t$ by ranking all candidates according to $d_k(x)$ in ascending order and selecting the n_k samples with the smallest discrepancy values to form $\mathcal{Q}_k^t$.

4.3 Consensus Enhanced Human-AI Collaboration

In this section, we detail how estimated consensus is leveraged at each client to enhance human-AI collaboration. Our idea is to treat the estimated consensus as a prior and first combine it with human annotations using a Bayesian framework, then fused with model predictions. In particular, we observe that when human experts have few historical annotations, the EDL model tends to produce small Dirichlet parameters, making the prior evidence insufficient to reliably reflect the consensus strength. To address this, we apply an EM algorithm at each client to calibrate the EDL outputs.

For the unlabeled data $\mathcal{D}_k^t$ arriving at client k, based on the query strategy mentioned above, $\mathcal{D}_k^t$ can be partitioned into two disjoint subsets: the queried set $\mathcal{Q}_k^t$ and its complement $\mathcal{D}_k^t \setminus \mathcal{Q}_k^t$. In the following, we detail how these two types of subsets are handled, respectively.

Fusion for Queried Samples. For samples $x \in \mathcal{Q}_k^t$, both the estimated consensus and human expert annotations are available. Recalling Assumption 1 that human expert consensus conforms to a latent Dirichlet distribution with parameter $\boldsymbol{\alpha}^\star(x)$, we treat the Dirichlet parameters $\hat{\boldsymbol{\beta}}_g(x)$ estimated by $g(x; w_g^{t-1})$ as a prior over the latent consensus distribution, and regard the expert annotation $h_k(x)$ as a categorical observation. Due to the conjugacy of the Dirichlet prior to the categorical likelihood, incorporating a single expert annotation admits a closed-form Bayesian update [Showalter *et al.*, 2024], in which the corresponding one-hot label is added directly to the Dirichlet parameters. In this way, we naturally fuse the two within a Bayesian framework.

[2]https://en.wikipedia.org/wiki/Kullback-Leibler_divergence

However, we observed that when the size of $\mathcal{U}_k^{\leq t}$ is limited, the EDL model tends to produce small Dirichlet parameters. Thus, we introduce an EM algorithm to calibrate the output of global EDL model $g(\cdot; w_g^{t-1})$, and the calibrated result at client k can be represented as $g(\cdot; w_g^{t-1}, z_k)$, where z_k is the calibration parameter. The above calibration procedure will be described in the next paragraph. Thus, the posterior Dirichlet parameters are finally updated as $\boldsymbol{\alpha'}(x) = \boldsymbol{\beta'}_g(x) + \text{one-hot}(h_k(x))$, where $\boldsymbol{\beta'}_g(x)$=$g(x; w_g^{t-1}, z_k)$.

Subsequently, we further fuse the $\boldsymbol{\alpha'}(x)$ with inference result of the task global model $f(x; \theta_{\text{g}}^{t-1})$. Specifically, we first compute the expected class probability $\hat{p}_c(x) = \frac{\alpha'_c}{\sum_{j=1}^{C} \alpha'_j}$ induced by $\boldsymbol{\alpha'}(x)$. Let $f_c(x; \theta_{\text{g}}^{t-1})$ denote the c-th entry of the global model predicted probability vector, and define the prediction confidence as $\lambda(x) = \max_c f_c(x; \theta_{\text{g}}^{t-1})$. We then weight the model prediction by its confidence as $f'_c(x; \theta_{\text{g}}^{t-1}) = \lambda(x)\, f_c(x; \theta_{\text{g}}^{t-1})$, and combine it with $\hat{p}_c(x)$ to obtain the fused probability and label as follows:

$$\tilde{p}_c(x) = \frac{f'_c(x; \theta_{\text{g}}^{t-1}) + \hat{p}_c(x)}{\sum_{j=1}^{C} \big(f'_j(x; \theta_{\text{g}}^{t-1}) + \hat{p}_j(x)\big)},\ y' = \arg\max_c \tilde{p}_c(x). \tag{5}$$

The fusion results of queried samples are denoted by $\Phi_1(\mathcal{I}_k^t, \mathcal{O}_k^t, \mathcal{U}_k^t) = \{(x, y') \mid x \in \mathcal{Q}_k^t\}$. For Φ_1, we do not apply confidence-threshold filtering because expert annotations are more likely to be consistent with the consensus in real-world and therefore should not be discarded.

Calibration of EDL Model. To calibrate the outputs of global model $\hat{\boldsymbol{\beta}}_g(x)$ at client k, we introduce a calibration parameter $z_k \in \mathbb{R}^+$ and define new Dirichlet parameters as $\boldsymbol{\beta'}_g(x) = [\beta'_1(x; z_k), ..., \beta'_C(x; z_k)]$, where $\beta'_c(x; z_k) = z_k e_c(x) + 1$. We employ an EM algorithm to estimate the parameter z_k. The EM algorithm[3] is an iterative method that alternates between estimating latent variables (E-step) and optimizing parameters (M-step) to maximize the likelihood with hidden variables. Specifically, for sample $x \in \mathcal{Q}_k^t$, we introduce a latent variable $\tilde{y}$ representing its fused label. The goal of the EM algorithm is to estimate the calibration parameter z_k by maximizing the following log-likelihood:

$$z_k^\star = \arg\max_{z_k > 0} \sum_{x \in \mathcal{Q}_k^t} \log \sum_{c=1}^{C} p(\tilde{y} = c, h_k(x), \mathbf{e}(x) \mid z_k)\,. \tag{6}$$

where $\mathbf{e}(x) = [e_1(x), \ldots, e_C(x)]$ denotes the evidence vector produced by the global EDL model for sample x.

Let $\mathbf{o}(x) = \text{one-hot}(h_k(x))$, $o_c(x)$ denotes the c-th component of the one-hot encoded vector. We directly present the formulas for the E-step and the M-step here, the detailed proof and implementation can be found in Appendix B.

- E-step: for sample $x \in \mathcal{Q}_k^t$, we estimate the posterior distribution of the $\tilde{y}$ under the current z_k during E-step, which can be expressed as:

$$q_c(x) = \frac{o_c(x) + (z_k e_c(x)) + 1}{\sum_{j=1}^{C} (o_j(x) + (z_k e_j(x)) + 1)}. \tag{7}$$

- M-step: Given the distribution $q_c(x)$ of the latent variable $\tilde{y}$, in M-step we update the calibration parameter z_k, we use z'_k to represent the updated z_k, then we have:

$$z'_k = \arg\max_{z_k > 0} \sum_{x \in \mathcal{Q}_k^t} \sum_{c=1}^{C} q_c(x) \cdot \log \frac{z_k e_c(x) + 1}{\sum_{j=1}^{C} (z_k e_j(x) + 1)}. \tag{8}$$

Fusion for Unqueried Samples. For samples $x \in \mathcal{D}_k^t \setminus \mathcal{Q}_k^t$, human annotations are unavailable. Therefore, their fusion relies solely on the model predictions $f(x; \theta_{\text{g}}^{t-1})$ and the calibrated estimated consensus $\boldsymbol{\beta'}_g(x) = g(x; w_g, z_k)$. Specifically, we first compute the expected class probability $\hat{p}_c(x) = \frac{\beta'_{g,c}(x; z_k)}{\sum_{j=1}^{C} \beta'_{g,j}(x; z_k)}$ induced by $\boldsymbol{\beta'}_g(x)$. Then use Eq.5 to compute the fused label y'. The confidence score s is obtained as $s = \max_c \tilde{p}_c(x)$. Finally, we apply a confidence threshold τ to filter unreliable labels. The fusion results for unqueried samples are denoted by $\Phi_2(\mathcal{I}_k^t, \mathcal{O}_k^t) = \{(x, y', s) \mid x \in \mathcal{D}_k^t \setminus \mathcal{Q}_k^t,\ s > \tau\}$.

In the end, we merge $\Phi_1(\mathcal{I}_k^t, \mathcal{O}_k^t, \mathcal{U}_k^t)$ and $\Phi_2(\mathcal{I}_k^t, \mathcal{O}_k^t)$ to construct the fusion set $\mathcal{M}_k^t$.

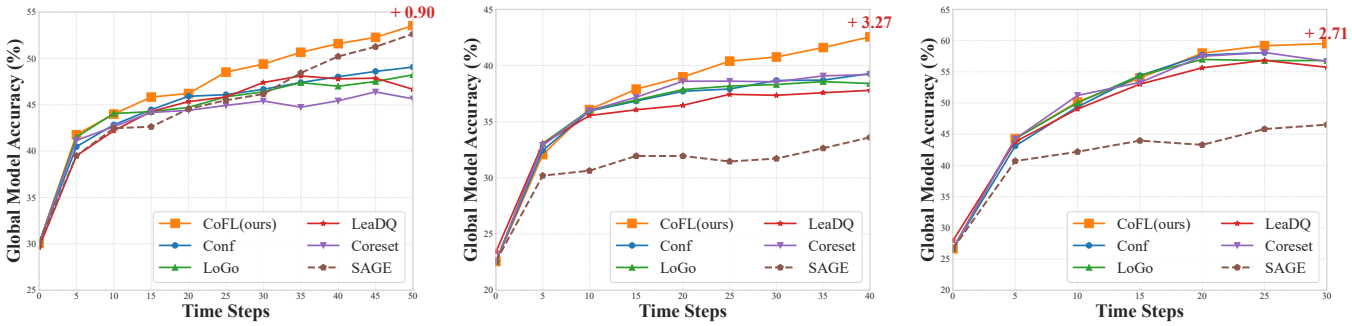
5 Experiments

5.1 Setup

FL System Setup. We simulate an FL system that learns a classification task from unlabeled streaming using a ResNet18 model[He *et al.*, 2016]. During each time step, each client performs $B = 10$ communication rounds and receives $|\mathcal{D}_k^t| = 50$ newly arrived samples. To model the Non-IID nature of client data, we use a Dirichlet distribution with parameter $\alpha = 0.5$ to assign different class proportions to different clients, which is a common setting in FL [Sun *et al.*, 2025].

Datasets. We evaluate CoFL on three datasets: CIFAR-10 [Krizhevsky *et al.*, 2009], CINIC-10 [Darlow *et al.*, 2018], and CIFAR-10H [Peterson *et al.*, 2019]. CIFAR-10 and CINIC-10 are widely used image classification benchmarks. To simulate the difference of human experts on these two datasets, we assign each expert h_k a unique confusion matrix $\mathbf{A}^k$, where $\mathbf{A}_{i,j}^k$ denotes the conditional probability $P(h_k(x) = i \mid y = j)$, with y being the ground-truth label of sample x. This process allows us to control and model diverse labeling behaviors across experts. The method of constructing a confusion matrix will be described in the next paragraph. CIFAR-10H is a real multi-expert annotated dataset that naturally captures differences among human annotators. For each sample, we compute the empirical label proportions to obtain a label probability distribution $\mathbf{p}(x) = [p_1(x), \ldots, p_C(x)]$. The human expert on each client then samples its label from $\mathbf{p}(x)$. To further emphasize subjective differences across experts, we introduce a temperature parameter $T_h = 4.0$ to smooth the distribution via $\tilde{p}_c(x; T_h) = \frac{p_c(x)^{1/T_h}}{\sum_{j=1}^{C} p_j(x)^{1/T_h}}$.

[3] https://en.wikipedia.org/wiki/Expectation-maximization_algorithm



(a) CIFAR-10 (b) CINIC-10 (c) CIFAR-10H

Figure 3: The accuracy of the global model at different time steps

Construction of Confusion Matrix. For each simulated expert h_m, we first sample a global ability level $A_m \sim \mathcal{N}(\mu_A, \sigma_A^2)$ to characterize the overall capability level of all experts in the FL system, and then sample a class specific bias vector $\boldsymbol{\delta}_m \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{cls}}^2 \mathbf{I})$ to model variations in performance across different classes. The probability of correctly labeling class j is defined as $p_{\text{corr},j}^{(m)} = \sigma(A_m + \delta_{m,j})$, where $\sigma(\cdot)$ is the sigmoid function. The confusion matrix is constructed column by column. For each true class, the diagonal entry is set as the probability of correct labeling, while the remaining probability mass is randomly distributed among the incorrect classes using a Dirichlet distribution, and each column is finally normalized to form a valid conditional probability distribution. We set $\mu_A = 1.0$, $\sigma_A = 1.0$, $\sigma_{\text{cls}} = 2.0$.

Baselines. To evaluate CoFL, we compare it against five baselines, including four FAL methods and a FSSL method. These baselines include: (1) Coreset [Sener and Savarese, 2017]: A representative diversity-based active learning method that selects a subset of samples that best represent the feature space. In our FL setting, this procedure is performed independently on each client's local data. (2) Conf: A representative uncertainty-based active learning method that selects samples with the lowest prediction confidence. Similarly, each client applies this strategy locally in the FL setting. (3) LoGo [Kim *et al.*, 2023]: An FAL approach that jointly considers local and global uncertainty to maintain both client-level and class-level diversity. (4) LeaDQ [Sun *et al.*, 2025]: An FAL framework that formulates the query strategy as a partially observable Markov decision process and solves it with multi-agent reinforcement learning. (5) SAGE [Liu *et al.*, 2025]: An FSSL method that leverages the discrepancy between local and global confidence to collaboratively generate and correct pseudo-labels.

Implementation Details. For the CIFAR-10, CINIC-10, and CIFAR-10H datasets, we set the query set size at each time step to $n_k = 10$, 20, and 30, respectively, and performed for 50, 40, and 30 time steps, respectively, to evaluate our method under different settings. We use CLIP ViT-B/32 [Radford *et al.*, 2021] to generate embeddings for the samples and employ a two-layer fully connected neural network as the EDL model, confidence threshold $\tau = 0.55$ and each client initializes with $|\mathcal{L}_k^0| = 100$ samples. More implementation details are provided in Appendix D.

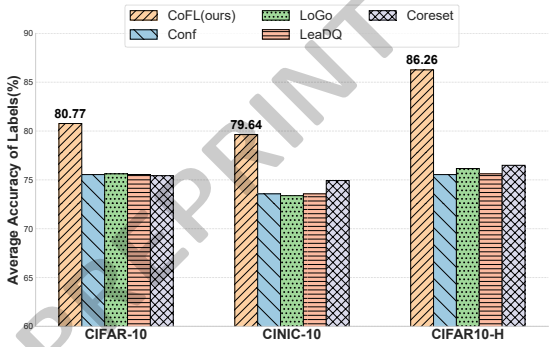


Figure 4: Average label accuracy of different methods across different datasets.

5.2 Main Results

Global Model Performance. Fig. 3 illustrates the performance of the global model over time steps for all methods across different datasets. We can observe that our proposed CoFL method consistently achieves the best performance across all three datasets at the final time step. On CIFAR-10H and CINIC-10 datasets, CoFL achieves around a 3% improvement over FAL methods, highlighting the advantage of considering the subjectivity of human experts. SAGE exhibits unstable performance across all datasets, which is likely because it heavily relies on the model performance. When the global model is limited in performance, SAGE cannot generate sufficiently reliable pseudo-labels and therefore fails to further improve performance.

Accuracy of Labels. Fig. 4 presents the average accuracy of labels for training samples during the local update at the final time step. We observe that the proposed CoFL method achieves the highest label accuracy across all three datasets. Specifically, on the CIFAR-10 dataset, CoFL reaches an accuracy of 80.77%, outperforming the suboptimal method LoGo, by a margin of 5.15%. This result further highlights the advantages of human-AI collaboration in generating more reliable labels. Note that since SAGE generates a large number of augmented samples that do not correspond with the original data, its label accuracy is not directly comparable to other methods. In addition, we visualize the label accuracy of the training samples at each client at the final time step. Detailed experimental results are provided in Appendix C.

5.3 Ablation Study

To evaluate the effectiveness of each module in CoFL, we conduct a comprehensive ablation study. Since the consensus estimation module serves as the foundation of human-AI collaboration and the query strategy, it cannot be removed in isolation. Therefore, we focus on the ablations of human-AI collaboration and the capability-aware sample querying strategy. For the ablation of the human-AI collaboration, we consider two settings: (1) directly training with annotations provided by human experts without any collaboration (denoted as w/o collaboration), and (2) without the EM-based calibration process for EDL (denoted as w/o EM). For the capability-aware sample querying strategy, we replace it with the Coreset method and evaluate the resulting performance gap to demonstrate its effectiveness (denoted as w/o capability-aware).

We report the ablation results in Table 1. On CIFAR-10, removing any module degrades performance, especially when using only human annotations, which reduces the accuracy from 53.5% to 43.8%, highlighting the importance of human-AI collaboration. For the EM-based calibration process, removing it leads to performance degradation on all datasets, indicating the importance of calibrating EDL. The capability-aware querying strategy is crucial on CIFAR-10 but less effective on CINIC-10 and CIFAR-10H. One possible explanation is that CINIC-10 has more diversity than CIFAR-10, the Coreset method may provide better coverage of the feature space. In contrast, the labels in CIFAR-10H are sampled from the same distributions of multiple human annotators, which makes the capabilities of different human experts more homogeneous. As a result, the capability-aware strategy cannot fully utilize its advantage.

Method	CIFAR-10	CINIC-10	CIFAR-10H
w/o collaboration	43.8	35.5	53.9
w/o EM	50.2	39.9	58.2
w/o capability-aware	50.1	**44.5**	**61.9**
CoFL	**53.5**	42.6	59.5

Table 1: Model accuracy (%) on different datasets in ablation study.

5.4 Experiments on Diverse Scenarios

To further investigate the adaptability of CoFL under different FL scenarios, we evaluate its performance with varying numbers of clients equipped with human experts and under different overall levels of human expert capability, i.e., the overall annotation accuracy when providing labels.

Varying Numbers of Human Expert Clients. Since SAGE is not influenced by human experts, we exclude it from this set of experiments. Given that there is no significant difference among all FAL methods in the main results, we select Coreset and LeaDQ for comparison with CoFL. We consider three settings with 4, 7, and 10 clients equipped with human experts, where 10 is the same as the setting used in the main results. For clients without experts, CoFL fuses estimated consensus with global model inference, while other methods train only on their initial labeled data.

We report the experimental results in Table 2. We observe that across all datasets and under all settings with different numbers of clients equipped with human experts, CoFL consistently achieves the best performance, demonstrating its strong adaptability.

Dataset	CIFAR-10			CINIC-10			CIFAR-10H		
Number	4	7	10	4	7	10	4	7	10
Coreset	42.5	43.1	45.6	34.5	35.6	39.2	49.4	55.4	56.7
LeaDQ	45.3	44.1	46.7	37.0	34.0	37.8	48.7	53.7	55.7
CoFL	**50.6**	**50.4**	**53.5**	**39.7**	**39.4**	**42.6**	**52.3**	**57.2**	**59.5**

Table 2: Model accuracy (%) comparison across different numbers of experts on three datasets.

Varying Human Expert Capability Levels. We evaluate different methods under varying expert capability levels, where higher levels represent more accurate annotations provided by all experts. For the CIFAR-10 and CINIC-10 datasets, we simulated the overall expert ability by adjusting the parameter μ_A when constructing the confusion matrix, where $\mu_A = 1.0$ for the medium level, $\mu_A = 0.2$ for the low level, and $\mu_A = 2.0$ for the high level. For the CIFAR-10H dataset, since expert annotations come from the same distribution, we simulate different levels of ability by smoothing the distribution with a temperature parameter T_h. We set $T_h = 2.0$, $T_h = 4.0$, and $T_h = 6.0$ to represent high, medium, and low ability, respectively.

We present the results in Table 3, where CoFL achieved the best performance in almost all cases, except for the high capability case on CIFAR-10H dataset. In addition, we observed that as the overall ability level of experts decreases, CoFL becomes more advantageous. For example, on CIFAR-10, the accuracy of Coreset and LeaDQ methods decreased by 14.4% and 16.6% respectively from high to low ability level, while CoFL only decreased by 7.9%. This indicates that the CoFL method can alleviate the negative impact of decreased capability, mainly because its consensus estimation can extract reliable signals from inconsistent annotations.

Dataset	CIFAR-10			CINIC-10			CIFAR-10H		
Level	Low	Med	High	Low	Med	High	Low	Med	High
Coreset	37.1	45.6	51.5	31.6	39.2	42.4	52.4	56.7	**62.7**
LeaDQ	38.5	46.7	55.1	30.4	37.8	43.6	50.8	55.7	60.5
CoFL	**49.5**	**53.5**	**57.4**	**36.2**	**42.6**	**45.4**	**58.3**	**59.5**	61.2

Table 3: Model accuracy (%) across different capability levels on different datasets.

6 Conclusions

In this paper, we propose CoFL, a human-AI collaborative FL method for addressing label scarcity in unlabeled streaming data. By leveraging human-AI complementarity and cross-client expert consensus, CoFL offers a new direction for learning from unlabeled data in FL. Future work can integrate CoFL with advanced FSSL and FAL methods for further improvement.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (No. 2024YFB4505502), the National Natural Science Foundation of China (No. 62372381) and the National Natural Science Foundation of China (No. 62302396).

References

[Cao *et al.*, 2023] Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, and Dacheng Tao. Knowledge-aware federated active learning with non-iid data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22279–22289, 2023.

[Chen *et al.*, 2024] Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11439–11449, 2024.

[Darlow *et al.*, 2018] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[Diao *et al.*, 2022] Enmao Diao, Jie Ding, and Vahid Tarokh. Semifl: Semi-supervised federated learning for unlabeled clients with alternate training. *Advances in Neural Information Processing Systems*, 35:17871–17884, 2022.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Jiang *et al.*, 2024] Zhida Jiang, Yang Xu, Hongli Xu, Zhiyuan Wang, and Chunming Qiao. Clients help clients: Alternating collaboration for semi-supervised federated learning. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1847–1860. IEEE, 2024.

[Kelly *et al.*, 2025] Markelle Kelly, Alex Boyd, Sam Showalter, Mark Steyvers, and Padhraic Smyth. Bayesian inference for correlated human experts and classifiers. *arXiv preprint arXiv:2506.05636*, 2025.

[Kerrigan *et al.*, 2021] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021.

[Kim *et al.*, 2023] SangMook Kim, Sangmin Bae, Hwanjun Song, and Se-Young Yun. Re-thinking federated active learning based on inter-class diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3944–3953, 2023.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Lee *et al.*, 2024] Seungjoo Lee, Thanh-Long V Le, Jaemin Shin, and Sung-Ju Lee. (fl)ˆ2: Overcoming few labels in federated semi-supervised learning. *Advances in Neural Information Processing Systems*, 37:43693–43714, 2024.

[Li *et al.*, 2022] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pages 965–978. IEEE, 2022.

[Liu *et al.*, 2025] Yijie Liu, Xinyi Shang, Yiqun Zhang, Yang Lu, Chen Gong, Jing-Hao Xue, and Hanzi Wang. Mind the gap: Confidence discrepancy can guide federated semi-supervised learning across pseudo-mismatch. *arXiv preprint arXiv:2503.13227*, 2025.

[Luo *et al.*, 2023] Jun Luo, Matias Mendieta, Chen Chen, and Shandong Wu. Pgfed: Personalize each client's global objective for federated learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3946–3956, 2023.

[Madras *et al.*, 2018] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

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

[Muttenthaler *et al.*, 2025] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

[Peterson *et al.*, 2019] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Raykar *et al.*, 2010] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca

Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

[Sener and Savarese, 2017] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Shao *et al.*, 2024] Jiawei Shao, Fangzhao Wu, and Jun Zhang. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. *Nature Communications*, 15(1):349, 2024.

[Showalter *et al.*, 2024] Samuel Showalter, Alex J Boyd, Padhraic Smyth, and Mark Steyvers. Bayesian online learning for consensus prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2539–2547. PMLR, 2024.

[Steyvers *et al.*, 2022] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022.

[Sun *et al.*, 2025] Yuchang Sun, Xinran Li, Tao Lin, and Jun Zhang. Learn how to query from unlabeled data streams in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20752–20760, 2025.

[Tang and Yu, 2025] Xiaoli Tang and Han Yu. Reputation-aware revenue allocation for auction-based federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20832–20840, 2025.

[Vaccaro *et al.*, 2024] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, 2024.

[Wang *et al.*, 2025a] Runze Wang, Jiahao Liu, Miao Hu, Yipeng Zhou, and Di Wu. Local differentially private release of infinite streams with temporal relevance. In *Proceedings of the ACM on Web Conference 2025*, pages 921–930, 2025.

[Wang *et al.*, 2025b] Zhihao Wang, He Bai, Wenke Huang, Duantengchuan Li, Jian Wang, and Bing Li. Federated recommendation with explicitly encoding item bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12792–12800, 2025.

[Wen *et al.*, 2023] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International Journal of Machine Learning and Cybernetics*, 14(2):513–535, 2023.

[Wu *et al.*, 2022] Xing Wu, Jie Pei, Cheng Chen, Yimin Zhu, Jianjia Wang, Quan Qian, Jian Zhang, Qun Sun, and Yike Guo. Federated active learning for multicenter collaborative disease diagnosis. *IEEE transactions on medical imaging*, 42(7):2068–2080, 2022.

[Yang *et al.*, 2025] Hang Yang, Zhiwu Li, and Witold Pedrycz. Adaptive deep learning from crowds. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 4263–4272, 2025.

[Yax *et al.*, 2024] Nicolas Yax, Hernán Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024.

[Yu *et al.*, 2021] Hongzheng Yu, Zekai Chen, Xiao Zhang, Xu Chen, Fuzhen Zhuang, Hui Xiong, and Xiuzhen Cheng. Fedhar: Semi-supervised online learning for personalized federated human activity recognition. *IEEE transactions on mobile computing*, 22(6):3318–3332, 2021.

[Zhang *et al.*, 2024] Zheng Zhang, Wenjie Ai, Kevin Wells, David Rosewarne, Thanh-Toan Do, and Gustavo Carneiro. Learning to complement and to defer to multiple users. In *European Conference on Computer Vision*, pages 144–162. Springer, 2024.

[Zhao *et al.*, 2023] Lei Zhao, Lin Cai, and Wu-Sheng Lu. Federated learning for data trading portfolio allocation with autonomous economic agents. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Zhao *et al.*, 2025a] Qing Zhao, Jielei Chu, Zhaoyu Li, Wei Huang, Zhipeng Luo, and Tianrui Li. Fedlgmatch: Federated semi-supervised learning via joint local and global pseudo labeling. *Knowledge-Based Systems*, page 113642, 2025.

[Zhao *et al.*, 2025b] Xuehan Zhao, Jiaqi Liu, Xin Zhang, Zhiwen Yu, and Bin Guo. Activehai: Active collection based human-ai diagnosis with limited expert predictions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 4282–4290, 2025.

[Zöller *et al.*, 2025] Nikolas Zöller, Julian Berger, Irving Lin, Nathan Fu, Jayanth Komarneni, Gioele Barabucci, Kyle Laskowski, Victor Shia, Benjamin Harack, Eugene A Chu, et al. Human–ai collectives most accurately diagnose clinical vignettes. *Proceedings of the National Academy of Sciences*, 122(24):e2426153122, 2025.