FLAIRE: ZERO-TRUST AND RARITY-AWARE CLIENT INCENTIVIZATION FOR HETEROGENOUS FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A challenge in Federated Learning is training under severe non-IID label skew. Incentivizing Rare-label clients is essential for maintaining global performance Existing incentive mechanisms often break down in practical deployments: they either assume honest client self-reporting of descriptors—exposing the system to strategic manipulation—or rely on server-side auxiliary datasets and external validators, violating core privacy and practical constraints. To resolve this dilemma, we propose **FLAIRE** (Federated Learning with Adaptive Incentives and Rewards Engine), an incentive architecture for heterogeneous FL built upon a **zero-trust**, **zero-auxiliary-data**, and **zero-metadata** foundation. FLAIRE achieves completely passive, metadata-free incentivization through two decoupled components: (1) a **client profiling engine** that infers local label distributions strictly via last-layer gradient inversion, requiring zero client-reported descriptors and zero server validation data; and (2) a **label-rarity-aware incentive mechanism** that assigns rewards inversely proportional to global label frequency. This design guarantees fair evaluation for rare-data providers ("Mavericks") while remaining resilient to free-riders and metadata tampering. FLAIRE is the first framework to unify passive, zero-trust profiling with rarity-aware contribution valuation in non-IID FL. Extensive evaluations on MNIST, Fashion-MNIST, EMNIST, and FEMNIST demonstrate that FLAIRE consistently boosts convergence speed and global model accuracy under extreme heterogeneity, achieving a minimum accuracy improvement of 5% across all settings and up to 10%–25% on the highly skewed benchmark scenarios tested.

1 INTRODUCTION

Federated learning (FL) McMahan et al. (2017); Bonawitz et al. (2019) is a distributed machine learning (ML) paradigm that enables multiple clients to collaboratively train a global model without sharing their private data. As depicted in Figure 1, each client trains a local model on its own data and shares only model updates with a central server, which processes them to produce the global model. This setup enables learning from decentralized datasets while preserving data privacy. The success of FL critically depends on the active participation of clients and the nature of the data they contribute Witt et al. (2023); Soltani et al. (2023); Tu et al. (2022). In scenarios where the trained model has commercial or societal value, it is important to quantify each client's contribution to the final model Gao et al. (2021). This is particularly important for clients with rare or under-represented data: although they may contribute relatively few samples, these samples can disproportionately improve performance on minority classes. Without appropriate incentives, such clients may withdraw from the training process, leading to degraded performance on under-represented classes.

Existing incentive mechanisms (IMs) in FL adopt a *performance-proportional* notion of fairness, where clients are rewarded according to the marginal improvement they have to the global model. In most works Xu et al. (2021); Salazar et al. (2026); Tastan et al. (2024) clients who are deemed to be more valuable receive a *higher-quality global model* in the *next training round*, under the assumption that their contributions justified greater influence. This creates a feedback effect in which the reduced quality model delivered each round further restricts an "underperforming" client's ability to contribute, causing them to receive increasingly weaker updates and compounding their perfor-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

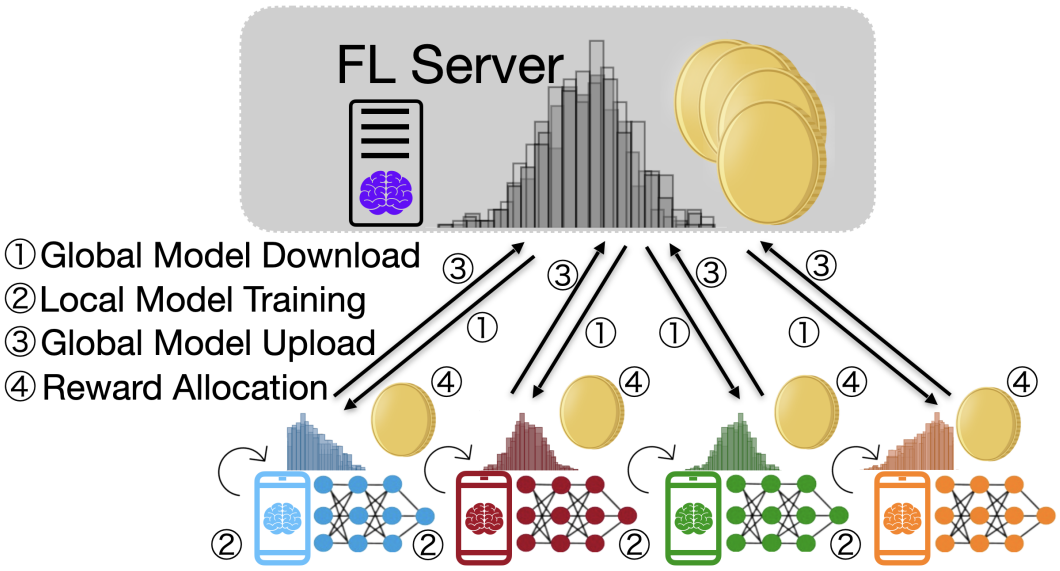


Figure 1: Federated Learning with Incentives and Rewards.

mance decline. While intuitive, this design implicitly assumes that clients have comparable and representative data distributions. In realistic scenarios however, where clients have highly non-IID label-limited datasets, they may appear to "under-perform" not because their data is low-quality, but because their data lacks overlap with test sets or contributes to under-represented classes that the global model struggles to learn. Conventional fairness-driven incentive schemes can unintentionally penalize honest participants whose data is structurally different rather than deficient.

Additionally, such approaches largely overlook the fact that, in practical federated settings, certain labels may be held by only a small number of clients. These clients are critical for learning under-represented regions of the label space, yet are systematically under-incentivized or misclassified as low contributors when rewards are computed without accounting for label scarcity. Prior mechanisms fail to explicitly encourage sustained participation from clients responsible for rare or structurally important labels.

To address these gaps, we propose FLAIRE, an incentive mechanism that extends current schemes by explicitly rewarding clients based on the *structural scarcity* and *representational importance* of the labels they contribute, rather than only on their immediate effect on aggregate accuracy. By accounting for local contributions to under-represented regions of the label space, FLAIRE encourages participation from clients whose data is essential but typically undervalued in existing methods. The objective is not merely to equalize perceived fairness, but to steer the federation toward improved *total model capability*, particularly for rare or difficult classes, by aligning incentives with long-term representational utility under non-IID data. To support this goal, we introduce a coverage-aware fairness metric that normalizes each client's performance against the maximum achievable accuracy given its own label coverage. This enables fairness to be evaluated relative to a client's feasible contribution domain, rather than through direct comparison across inherently unequal tasks. By separating structural disadvantage from algorithmic unfairness, FLAIRE provides a more principled framework for incentive design and fairness assessment in heterogeneous, realistic FL systems.

Concretely, our contributions are:

- A framework for quantifying client contributions with a focus on underrepresented labels.
- An incentive mechanism that rewards clients proportionally to both their contribution and the rarity of their data.
- A fairness definition that works on highly label-skewed and high label-scarcity scenarios.
- A trustless mechanism that does not rely on self-reported metrics or external information about client data.

- A general contribution evaluation mechanism that does not rely on validation from external datasets.
- Experiments demonstrating the effectiveness of the proposed mechanism on diverse non-IID settings on FL.
- Efficiency on multi-class label datasets.

Experimental results demonstrate that FLAIRE satisfies all of the above desiderata while achieving a minimum performance improvement of 5% across all evaluated datasets with experiments reaching 10% and even 25%, on diverse settings of benchmark datasets. These gains are obtained without any trust assumptions, or reliance on external datasets. These improvements address a fundamental need in large multi-class settings with skewed and label-scarce distributions.

2 Backround & Related Work

Incentive Mechanisms and Contribution Evaluation. In FL, clients must be motivated to participate consistently, especially when their data is valuable but scarce. IMs address this challenge by rewarding clients according to their impact on the global model. An effective IM has two key components: (i) a method for evaluating each client's contribution and (ii) a protocol for fair reward distribution based on these evaluations. Central to this is *contribution evaluation* (CE), the process of quantifying each client's influence on the final model. CE is therefore the foundation upon which fair and reliable IMs are built. However, *reward allocation* must also account for the inherent heterogeneity of client data arising from non-IID distributions. Figure 1 showcases the design of an FL system with an IM and client reward allocation.

Contribution Evaluation in FL. In conventional ML, data valuation is typically performed by evaluating the model utility of a trained model on a centralized test dataset. Since all training data is centrally available, the contribution of individual data points can be assessed directly using standard performance metrics. However, in FL, CE methods must operate under privacy constraints, making the evaluation of client contributions more complex. Existing approaches can be categorized as: (i) data-level evaluation, (ii) model utility-based evaluation, (iii) Shapley-value (SV) analysis, and (iv) statistical metrics Soltani et al. (2023). Absolute metrics can be computed independently for each client, e.g., by measuring the difference in loss between a local and global model,while relative metrics, require comparisons across clients or model updates. Initial works on absolute metrics in FL focus on accuracy or dataset size McMahan et al. (2017); Feng et al. (2019). Another cohort of methods consider the importance of labels in CE by evaluating the alignment of client and server labels before training and Ma et al. (2021); Li et al. (2021). These methods however require honest reporting of client metrics, which is a significant unrealistic assumption for clients in real-world FL settings. A large cohort of methods utilize a server-side validation dataset to evaluate all clients Zhang et al. (2021); Deng et al. (2021). The origin of such a dataset —which has to be class-balanced by design for valid results— is overlooked. In statistical-metric relative methods, where Zhao et al. Zhao et al. (2023) proposed a method based on the Euclidean distance between client model updates, while Witt et al. Witt et al. (2024) introduced a peer-to-peer truth serum approach that evaluates contributions by correlating client and global performance on a public dataset. Fall short as distance metrics and peer agreement across non-iid clients are misleading.

Shapley Values. Recently most efforts have shifted towards the adoption of SVs in FL. Wang et al. Wang et al. (2020) introduce fedSV, a federated variant of the SV. SV approximations later were proposed like TMC-SHAPLEY Ghorbani & Zou (2019) and GT-Shapley Jia et al. (2019), as SV evaluation requires factorial evaluations. Xu et al. Xu et al. (2021) further extend the fedSV approach to the class-specific setting, enabling the estimation of each client's contribution per class. However they both still implicitly assume an IID setting without different label counts across clients.

A significant CE method proposed by by Tastan et al. Tastan et al. (2024) acknowledges the need for label-specific CE in settings where clients have heterogenous label distributions by using a class-specific SV (CSSV). However, by averaging the calculated cosine similarities, they implicitly assume that every client contributes to each label. As a result, clients that own only a subset of the labels—especially those providing rare or unique classes—are unfairly penalized even when providing high-quality models, since their absence of gradients for missing labels is treated as zero

Feature	FLAIRE (Ours)	Wang et al. & Xu et al.	ShapFed	IMFL
No private information sharing about the data	●	●	●	○
No balanced auxiliary datasets used	●	●	●	○
Efficient contribution evaluation metric	●	◐	◐	○
Efficient use for multiple class datasets	●	◐	◐	○
Label Skew Robustness	●	○	○	◐
Label Rarity Support	●	○	○	◐

● Fully supported ◐ Partially supported ○ Not supported

Table 1: Comparison of FLAIRE with existing works

contribution. This bias skews the aggregation process toward clients with broader label coverage, even if their model updates are of lower quality.

Huang et al. Huang et al. (2024) propose IMFL-AIGC, an IM for FL empowered by artificial-intelligence-generated content (AIGC). They aim to mitigate the label unavailability in non-IID data distributions using synthetic data generated by a generative model. While this approach provides an empirically validated mechanism under realistic data heterogeneity assumptions, its reliance on AIGC introduces practical limitations. Generating synthetic data that faithfully represents each client's domain is computationally intensive, and the quality of generative models can degrade substantially when the generative distribution does not align with the task-specific data distribution. Consequently, such methods may struggle in federated settings where generative models fail to capture rare or domain-specific features.

Finally, these methods show significant computational overhead and are evaluated primarily on datasets with a small number of labels (e.g., 2, 4, 8, or 10). Evaluations on larger, real-world datasets and more realistic deployment settings remain largely unexplored. This limitation is particularly important in the context of AI-generated content, where the computational cost of scaling such approaches to large datasets becomes prohibitively expensive.

Comparison with FLAIRE. In contrast to the above approaches, **FLAIRE** introduces a *principled*, *scalable* and *efficient* IM targeting *full label coverage* of the training task and *fair label representation* on *label-skewed* and *label-scarce* settings without reliance on auxiliary balanced data that is either shared or generated. Our proposed IM infers client label distributions directly and incentivizes participation from clients holding minority labels, aiming to maximize label coverage and achieve a more comprehensive global model without relying on synthetic data generation. FLAIRE's fairness goal is to uncover the underlying data structure of participants and use them for evaluation.

3 Incentive Mechanism Design

Problem formulation. We consider a federated learning (FL) system comprising a set of clients $\mathcal{C} = \{1, 2, \ldots, N\}$ collaboratively training a global model parameterized by $\mathbf{w}$ under the coordination of a central server. Each client $i \in \mathcal{C}$ holds a local dataset $\mathcal{D}_i = \{\xi_{i,j}\}_{j=1}^{n_i}$, where each input sample $\xi_{i,j} = (\mathbf{x}_{i,j}, y_{i,j})$ consists of a feature $\mathbf{x}_{i,j}$ and its corresponding label $y_{i,j}$. From the total set of available classes $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$ defining the global training task, each client's dataset $\mathcal{D}_i$ contains samples associated with a subset $\mathcal{L}_i \subseteq \mathcal{L}$ of the label space, where $|\mathcal{L}_i| \leq K$. The local training objective of client i is defined as

$$F_i(\mathbf{w}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\big[\ell(f_{\mathbf{w}}(\xi_i))\big], \tag{1}$$

and the overall global optimization problem is expressed as

$$\min_{\mathbf{w}} F(\mathbf{w}) = \sum_{i=1}^{N} \frac{n_i}{n} F_i(\mathbf{w}), \quad \text{where } n = \sum_{i=1}^{N} n_i. \tag{2}$$

Clients participate in each round of the FL process with probability $\boldsymbol{p}_i^t \in [0,1]^N$. The goal is to given each cleint's label distribution $\mathcal{L}_i$ to calculate a reward $\boldsymbol{R}_i^t \in \mathbb{R}^N$ based on clients' performance and label-importance to the task.

3.1 Proposed Protocol - Solution

Client Profiling in Non-IID FL. To address the challenges posed by heterogeneous and label-skewed data distributions in federated learning (FL), we introduce a *client profiling* step that estimates each client's label distribution without accessing its private data. All clients are initially treated as anonymous participants with unknown label subsets $\mathcal{L}_i \subseteq \mathcal{L}$. After the initial epoch of local training, each client transmits the resulting gradient updates $\Delta\omega_i^1$ back to the server. The server then uses a gradient inversion method Yin et al. (2021); Zhao et al. (2020) to reconstruct approximate label information from these updates, producing an estimated label distribution $\hat{p}_i(l)$ for each client (Alg. 1).

Algorithm 1 Federated Learning with Grad-based Profiling

Initialize: Global model w_0, learning rate η, momentum μ, profiling interval rounds r, total rounds T
for each profiling round $t = 1, 2, \ldots, \leq r \leq T$ **do**
  for each client $k \in C_t'$ **in parallel do**
    $w_k^1 \leftarrow \text{ClientLocalTrain}(w_0, D_k^{train})$
    $\Delta_k^1 \leftarrow w_k^1 - w_0$ {Send model delta to server}
  end for
  Server-side profiling:
  for each client $k \in C_t$ **do**
    $\widehat{g}_k^t \leftarrow \text{EffLR}(\Delta_k^t, \eta, \mu)\ \widehat{g}_k^t \leftarrow -\frac{1}{\eta}\Delta_k^t$
    Extract final-layer gradients $(\widehat{g}_{k,w}^t, \widehat{g}_{k,b}^t)$
    $\hat{p}_i(l) \leftarrow \text{GradInv}(\widehat{g}_{k,w}^t)$ iDLG$(\widehat{g}_{k,w}^t)$
  end for
  Server aggregates: $w_{t+1} \leftarrow \sum_{k \in C_t'} \frac{n_k}{N} w_k^t$
end for

Label Rarity incentive Incentive Protocol. We design an incentive mechanism that rewards clients according to the rarity of the labels they contribute. Let C denote the set of clients and $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$ the set of labels. Each client $i \in C$ possesses a label set $\mathcal{L}_i \subseteq \mathcal{L}$. For each label $\mathcal{L}_i \subseteq \mathcal{L}$, we compute its client frequency and define a per-label payoff with $L > 0$ as a base reward constant inversely proportional:

$$f_y = \big|\{\, i \in C \mid y \in L_i \,\}\big| \quad \lambda_y = \frac{L}{f_y},$$

Client Reward. Let $C_i(l)$: marginal contribution (e.g. per-label Shapley approximation) of client i on label l. If client i does not train on label l, then $C_i(l) = 0$. We use the CGSV_i which denotes the cosine similarity-based SV contribution approximation for client i, as shown by Xu et al. (2021); Jiang et al. (2023) capturing how aligned their contribution is with the aggregate Shapley vector. Let λ_i denote the vector of the rarity of the labels λ_y possessed by client i, defined as the inverse frequency of each label across all clients. The final incentive for client i is computed as their dot product:

$$I_i = \lambda_i^{\alpha} \cdot C_i^{\beta}$$

where $\alpha, \beta > 0$ are tunable parameters that control the relative importance of contribution quality versus label rarity. The multiplicative formulation ensures that clients are rewarded most when they provide *both valuable and rare contributions* with per-label granularity for all clients.

Normalized contributions for partial-label clients. We assume *honest* clients who train on a subset of the total label space. To fairly evaluate each client's contribution, we normalize their per-label contributions by the maximum achievable contribution given their label subset. Formally, for a client i training on labels $L_i \subseteq \{1, \ldots, L\}$ with per-label contributions $C_i(l)$, we define the maximum achievable contribution as

$$C_i^{\max} = \sum_{l \in L_i} C_i^{\max}(l),$$

where $C_i^{\max}(l)$ is the theoretical maximum contribution on label l. The normalized contribution for each label is then

$$\hat{C}_i(l) = \frac{C_i(l)}{C_i^{\max}}, \quad l \in L_i.$$

This normalization ensures that clients are evaluated relative to what they could possibly contribute, avoiding penalization for partial label coverage while maintaining comparability across clients.

3.2 Reward construction combining marginal contribution and label rarity (Alg. 2)

Label rarity weight (soft inverse-frequency). For each label l define an un-normalized rarity score

$$f_l = \sum_{i=1}^{N} \mathbb{1}_{\{\hat{p}_i(l)>0\}}, \quad \forall l \in \mathcal{L}, \quad \forall l \in \mathcal{L} = \bigcup_{i=1}^{N} L_i \tag{3}$$

Label Rarity Weights. Calculating the corresponding weights

$$w(l) \;=\; \big(f(l)^{-1}\big)^{\alpha} \;=\; \big(1/f(l)\big)^{\alpha}, \quad \forall l \in \mathcal{L} \tag{4}$$

If $\alpha = 0$ the scheme reduces to pure contribution-based rewards (no rarity). If $\alpha = 1$ labels are weighted strictly by inverse frequency.

Per-label Maximum Contribution (Shapley Ceiling). The Best observed Shapley contributor for label l in this round is selected as the reference ceiling.

$$C^{\max}(l) = \max_{i: l \in L_i} C_{i,l}, \quad \forall l \in \mathcal{L} \tag{5}$$

Per-client Maximum Achievable Contribution. The maximum amount that client C_i can contribute if they were best-in-class on every label they own.

$$C_i^{\max} = \sum_{l \in L_i} C^{\max}(l), \quad \forall i = 1, \ldots, N \tag{6}$$

Normalized Per-label Contribution. The client contribution is then normalized in terms of the maximum possible client contribution

$$\hat{C}_i(l) = \frac{C_{i,l}}{C_i^{\max}}, \quad \forall l \in L_i \tag{7}$$

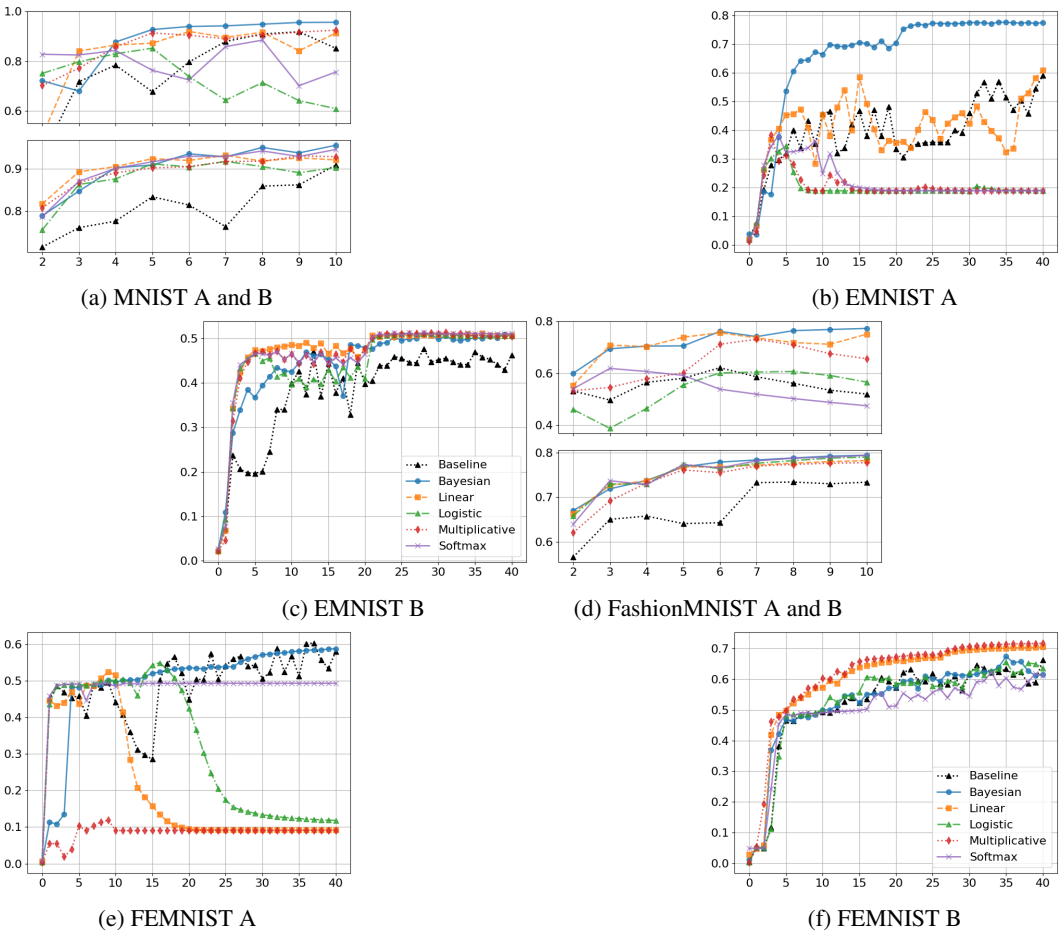
Raw Client Reward (Weighted by Label Rarity). The raw reward is calculated by summarizing over the product of the label rarity weight and the normalized client contribution per label.

$$R_i^{\text{raw}} = \sum_{l \in L_i} w(l)\, \hat{C}_i(l) = \frac{1}{C_i^{\max}} \sum_{l \in L_i} w(l)\, C_{i,l} \tag{8}$$

Normalize to Budget. Finally, normalize rewards so they sum to the available budget B (or to 1 for relative scores). If B is the total payout available, the base weighted-score for client i is

$$R_i \;=\; B \cdot \frac{\tilde{R}_i}{\sum_{j=1}^{N} \tilde{R}_j + \delta}, \tag{9}$$

with small $\delta > 0$ to avoid division by zero.



(a) MNIST A and B

(b) EMNIST A

(c) EMNIST B

(d) FashionMNIST A and B

(e) FEMNIST A

(f) FEMNIST B

Figure 2: FLAIRE performance across MNIST, EMNIST, Fashion-MNIST and FEMNIST (Accuracy per Round)

4 Experiments and Results

Datasets. Experiments include four benchmark FL datasets: MNIST Lecun et al. (1998), FMNIST Xiao et al. (2017), EMNIST Cohen et al. (2017) and FEMNIST Caldas et al. (2019). There are two 10-label datasets, then EMNIST with 47 labels and FEMNIST with 62 labels. Similar non-IID setups have been introduced in Non-IID client FL Zhao et al. (2018), in Wu et al. (2024) for personalized FL and in FairFed Liu et al. (2024).

In FairFed, non-IID data partitions are defined by label-group coverage rather than per-class isolation. Specifically, the Two-Group setting consists of two equally sized client groups, each restricted to disjoint subsets of five labels (0–4 and 5–9). The Three-Group setting partitions clients into three equal groups covering label subsets of sizes four (0–3), three (4–6), and three (7–9), respectively. Their Four-Group setting again splits labels into two or three per client with unequal client group sizes. IMFL-AIGL initially assumes an overlapping non-IID setting, where participants' data distributions partially intersect across classes. It uses clients with local datasets containing varying numbers of labels. However, they do not train under this setting, as they augment the data by filling in missing classes; otherwise, clients with one or two labels, which collapses training.

Client Groups. We extend the experimental methodologies and generalized them for arbitrary client groups and training subsets. In our experimental setup, clients are organized into a small number of groups, with each dataset partitioned into a number of subgroups depending on its label space. Within each group, clients are restricted to samples from a subset of the datasets' labels, resulting in highly non-IID local datasets. Combining the datasets of all the groups involved yields the

Table 2: MNIST and FMNIST after 10 training rounds

Method	MNIST A	MNIST B	FMNIST A	FMNIST B
Baseline	85.15%	90.08%	51.99%	73.35%
Bayesian	95.59% (10.44%)	95.48% (5.39%)	77.25% (25.26%)	79.39% (6.04%)
Linear	91.23% (5.98%)	92.08% (2.00%)	75.03% (23.04%)	78.19% (4.84%)
Logistic	60.89% (-24.26%)	90.21% (0.13%)	56.66% (-20.59%)	78.98% (5.64%)
Multipli.	91.41% (6.26%)	92.82% (2.74%)	65.61% (13.62%)	77.72% (4.38%)
Softmax	75.60% (-9.55%)	94.48% (0.43%)	47.48% (-4.51%)	79.38% (6.04%)

full original label space. Cross-group clients do not share any of the labels so the client groups in this scenario are non-overlapping (NOP). The dataset that belongs to each group is divided equally among participating clients. Thus, clients that belong in the same group share a common label space but don't share any samples, and groups with more participants have less samples per client.

MNIST and Fashion-MNIST. We select a minimum of three to four labels per group to guarantee stable training: when clients have fewer than three labels, we observe that local training is significantly degraded to the point that it becomes unstable and global training does not converge. We divide the ten digits into three groups: Group 1 contains digits {0,1,2}, Group 2 contains {3,4,5}, and Group 3 contains {6,7,8,9}. In *Setting A* we set up a total of $n = 10$ clients with 3 in the first group, 3 in the second group and 4 in the final group. In *Setting B* we set up a total of $n = 24$ participants, and groups are split into 12 clients for the first group, 8 clients for the second and 4 clients for the last one.

EMNIST and FEMNIST. Evaluating client incentive schemes on large, multi-class datasets remains a significant challenge, as most existing schemes are typically assessed on datasets with only 10 classes. For **EMNIST**, we partition the dataset into five groups: Group 1 and Group 5 each contain 10 labels, while Groups 2, 3, and 4 each contain 9 labels. In *Setting A*, each group is represented by a single client. In *Setting B*, Groups 1 and 2 each consist of 8 clients, Group 3 consists of 4 clients, and Groups 4 and 5 each consist of 2 clients, resulting in a total of 20 participants. For **FEMNIST**, the 62 labels are divided into four groups: Group 1 contains 11 labels, Groups 2 and 3 each contain 20 labels, and Group 4 contains the remaining 11 labels. As with EMNIST, *Setting A* assigns a single client to each group. In *Setting B*, Groups 1 and 2 each consist of 8 clients, Group 3 consists of 4 clients, and Group 4 consists of 2 clients, yielding a total of 20 participants. To the best of our knowledge, existing work does not explore incentive mechanisms under conditions of severe label heterogeneity and label scarcity, especially in multi-label datasets such as EMNIST and FEMNIST Mukhtiar et al. (2025). Setting A for both these large multi-label datasets is highly sensitive to training, as data subsets are held by single clients.

Experimental setup. We simulate the above settings by initializing each client with a base participation probability of 0.5, representing the likelihood that a client participates after accepting its reward. In each round, client participation is determined via a Bernoulli trial, with the participation probability updated based on the cumulative contribution score generated by the incentive mechanism. This score is passed through one of five activation functions, which modulate each client's participation probability across rounds. The baseline setting refers to unmodified probabilities across

Table 3: EMNIST and FEMNIST after 40 training rounds

Method	EMNIST A	EMNIST B	FEMNIST A	FEMNIST B
Baseline	59.00%	45.17%	58.06%	66.28%
Bayesian	77.49% (18.49%)	50.64% (5.47%)	58.80% (-0.74%)	61.44% (-4.84%)
Linear	60.79% (16.70%)	50.68% (5.51%)	09.10% (-49.06%)	70.43% (4.15%)
Logistic	18.95% (-40.05%)	50.58% (5.61%)	11.12% (-46.95%)	63.54% (-2.74%)
Multipli.	18.90% (-40.10%)	50.42% (5.25%)	09.10% (-49.06%)	71.68% (5.40%)
Softmax	18.95% (-40.05%)	51.11% (5.94%)	49.35% (-8.71%)	61.07% (-5.21%)

rounds. We are using 5 different activation functions, *Bayesian*, *Linear*, *Logistic*, *Multiplicative* and *Softmax*. Accuracy traces for all datasets are reported in Figure 2. Final round centralized accuracies for MNIST and Fashion-MNIST are depicted in Figure 2, while for EMNIST and FEMNIST in Figure 3. For all experiments we evaluate using FedAvg with different activation functions on the datasets and settings discussed[1].

Results. Across all datasets and evaluation settings, the Bayesian activation function consistently outperforms the baseline, yielding an improvement of at least 5%. On MNIST and FMNIST, Bayesian activation achieves the largest gains overall, with improvements of up to +10.44% and +25.26%, respectively. This indicates that uncertainty-aware incentive scaling provides robust benefits during early training. Linear and multiplicative activations demonstrate stable, though smaller, improvements, particularly on simpler datasets, suggesting that smooth, monotonic incentive mappings can positively influence convergence. However, these gains are less consistent on more complex datasets, where performance becomes sensitive to activation choice. In contrast, logistic and softmax activations exhibit unstable behavior, especially in setting A, where accuracy drops of up to -24.26% on MNIST and -40.05% on EMNIST are observed. While these functions occasionally preserve or slightly improve performance in setting B, their sharp non-linearities appear to disrupt convergence, particularly under heterogeneous data distributions.

CONCLUSION

This work proposed FLAIRE, an adaptive incentive mechanism for federated learning that mitigates the effects of non-IID data to improve task coverage and accelerate convergence. Our contributions include a granular, importance-aware framework that models client profiles in depth and explicitly addresses fairness under heterogeneous data distributions. By incorporating label-rarity and Shapley-based contribution normalization, FLAIRE encourages clients to provide high-quality models with data that is both scarce and valuable to the global model. This principled reward structure promotes equitable participation across heterogeneous clients while improving robustness and convergence in highly non-IID federated settings. Overall, our results indicate that incentive mechanism design must be adaptive to the task and its underlying heterogeneity: Shapley values capture contribution quality, but are insufficient without explicit mechanisms for rewarding rare and informative labels.

[1]The code will be released upon acceptance.

REFERENCES

Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. Towards federated learning at scale: System design. In A. Talwalkar, V. Smith, and M. Zaharia (eds.), *Proceedings of Machine Learning and Systems*, volume 1, pp. 374–388, 2019. URL https://proceedings.mlsys.org/paper_files/paper/2019/file/7b770da633baf74895be22a8807f1a8f-Paper.pdf.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2019. URL https://arxiv.org/abs/1812.01097.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017. URL https://arxiv.org/abs/1702.05373.

Yongheng Deng, Feng Lyu, Ju Ren, Yi-Chao Chen, Peng Yang, Yuezhi Zhou, and Yaoxue Zhang. Fair: Quality-aware federated learning with precise user incentive and model aggregation. *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pp. 1–10, 2021. doi: 10.1109/INFOCOM42981.2021.9488743.

Shaohan Feng, Dusit Niyato, Ping Wang, Dong In Kim, and Ying-Chang Liang. Joint Service Pricing and Cooperative Relay Communication for Federated Learning . In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pp. 815–820, Los Alamitos, CA, USA, July 2019. IEEE Computer Society. doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00148. URL https://doi.ieeecomputersociety.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00148.

Liang Gao, Li Li, Yingwen Chen, Wenli Zheng, ChengZhong Xu, and Ming Xu. Fifl: A fair incentive mechanism for federated learning. ICPP '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450390682. doi: 10.1145/3472456.3472469. URL https://doi.org/10.1145/3472456.3472469.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. *Proceedings of the 36th International Conference on Machine Learning*, 97:2242–2251, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Guangjing Huang, Qiong Wu, Jingyi Li, and Xu Chen. Imfl-aigc: Incentive mechanism design for federated learning empowered by artificial intelligence generated content. *IEEE Transactions on Mobile Computing*, 23(12):12603–12620, 2024. doi: 10.1109/TMC.2024.3414138.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 89:1167–1176, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/jia19a.html.

Meirui Jiang, Holger R. Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair Federated Medical Image Segmentation via Client Contribution Estimation . *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16302–16311, June 2023. doi: 10.1109/CVPR52729.2023.01564. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01564.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Anran Li, Lan Zhang, Juntao Tan, Yaxuan Qin, Junhao Wang, and Xiang-Yang Li. Sample-level data selection for federated learning. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pp. 1–10, 2021. doi: 10.1109/INFOCOM42981.2021.9488723.

Yiqi Liu, Shan Chang, Ye Liu, Bo Li, and Cong Wang. Fairfed: Improving fairness and efficiency of contribution evaluation in federated learning via cooperative shapley value. pp. 621–630, 2024. doi: 10.1109/INFOCOM52122.2024.10621438.

Jiahua Ma, Xinghua Sun, Wenchao Xia, Xijun Wang, Xiang Chen, and Hongbo Zhu. Client selection based on label quantity information for federated learning. In *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, 2021. doi: 10.1109/PIMRC50174.2021.9569487.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Noorain Mukhtiar, Adnan Mahmood, Yipeng Zhou, Jian Yang, Jing Teng, and Quan Z. Sheng. Federated learning at the forefront of fairness: A multifaceted perspective. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25), Survey Track*, pp. 10603–10611. IJCAI, 2025. doi: 10.24963/ijcai.2025/1177. URL https://www.ijcai.org/proceedings/2025/1177.pdf.

Teresa Salazar, Helder Araujo, Alberto Cano, and Pedro Henriques Abreu. A survey on group fairness in federated learning: Challenges, taxonomy of solutions and directions for future research. *Artificial Intelligence Review*, 2026. ISSN 1573-7462. doi: 10.1007/s10462-025-11475-5. URL https://doi.org/10.1007/s10462-025-11475-5.

Behnaz Soltani, Yipeng Zhou, Venus Haghighi, and John C. S. Lui. A survey of federated evaluation in federated learning. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 6769–6777. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/758. URL https://doi.org/10.24963/ijcai.2023/758. Survey Track.

Nurbek Tastan, Samar Fares, Toluwani Aremu, Samuel Horváth, and Karthik Nandakumar. Redefining contributions: Shapley-driven federated learning. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 5009–5017, 8 2024. doi: 10.24963/ijcai.2024/554. URL https://doi.org/10.24963/ijcai.2024/554. Main Track.

Xuezhen Tu, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Yang Zhang, and Juan Li. Incentive mechanisms for federated learning: From economic and game theoretic perspective. *IEEE Transactions on Cognitive Communications and Networking*, 8(3):1566–1593, 2022. doi: 10.1109/TCCN.2022.3177522.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. *A Principled Approach to Data Valuation for Federated Learning*, pp. 153–167. Springer International Publishing, Cham, 2020. ISBN 978-3-030-63076-8. doi: 10.1007/978-3-030-63076-8_11. URL https://doi.org/10.1007/978-3-030-63076-8_11.

Leon Witt, Mathis Heyer, Kentaroh Toyoda, Wojciech Samek, and Dan Li. Decentral and incentivized federated learning frameworks: A systematic literature review. *IEEE Internet of Things Journal*, 10(4):3642–3663, 2023. doi: 10.1109/JIOT.2022.3231363.

Leon Witt, Usama Zafar, KuoYeh Shen, Felix Sattler, Dan Li, Songtao Wang, and Wojciech Samek. Decentralized and incentivized federated learning: A blockchain-enabled framework utilising compressed soft-labels and peer consistency. *IEEE Transactions on Services Computing*, 17(4): 1449–1464, 2024. doi: 10.1109/TSC.2023.3336980.

Leijie Wu, Song Guo, Yaohong Ding, Junxiao Wang, Wenchao Xu, Yufeng Zhan, and Anne-Marie Kermarrec. Rethinking personalized client collaboration in federated learning. *IEEE Transactions on Mobile Computing*, 23(12):11227–11239, 2024. doi: 10.1109/TMC.2024.3396218.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16104–16117, 2021.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16332–16341, 2021. doi: 10.1109/CVPR46437.2021.01607.

Zhebin Zhang, Dajie Dong, Yuhang Ma, Yilong Ying, Dawei Jiang, Ke Chen, Lidan Shou, and Gang Chen. Refiner: A reliable incentive-driven federated learning system powered by blockchain. *Proceedings of the VLDB Endowment*, 14(12):2659–2662, 2021.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020. URL https://arxiv.org/abs/2001.02610.

Jianxin Zhao, Xinyu Chang, Yanhao Feng, Chi Harold Liu, and Ningbo Liu. Participant selection for federated learning with heterogeneous data in intelligent transport system. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):1106–1115, 2023. doi: 10.1109/TITS.2022.3149753.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. URL https://arxiv.org/abs/1806.00582.

A Appendix

Theorem 1 (Label-Space Fairness)**.** *Let R_i be the reward defined above with $\alpha \in [0, 1]$ and $\beta \geq 0$. Then the following properties hold:*

1. ***No inflation for zero contribution.*** *If $C_i(l) = 0$ for all l, then $R_i = 0$ for any α, β.*
2. ***Bounded rarity influence.*** *For any two labels l_1, l_2 with frequencies $f(l_1)$ and $f(l_2)$, their relative influence satisfies*
$$\frac{\tilde{w}(l_1)}{\tilde{w}(l_2)} = \left(\frac{f(l_2)}{f(l_1)}\right)^{\alpha},$$
ensuring label-rarity scales contributions smoothly and cannot dominate marginal contribution when $\alpha < 1$.
3. ***Fairness for disjoint label supports.*** *If two clients i and j train on disjoint label sets, their reward ratio is*
$$\frac{R_i}{R_j} = \frac{\sum_{l \in L_i} \tilde{w}(l)\hat{C}_i(l)}{\sum_{l \in L_j} \tilde{w}(l)\hat{C}_j(l)}$$
meaning clients are compared only on labels they actually train and are not penalized for lacking global label coverage.

A.1 Formal Properties of Client Rewards

The reward in Definition 1 is intentionally simple and interpretable. The following theorem formalizes four standard desiderata (budget-balance, null-player, symmetry, monotonicity) adapted to the partial-label setting.

Theorem 2 (Basic axiomatic properties)**.** *Under Definition 1 the reward allocation $R = (R_1, \ldots, R_N)$ satisfies:*

1. ***(Budget-balance)*** $\sum_{i=1}^{N} R_i = B \cdot \dfrac{\sum_i \widetilde{R}_i}{\sum_j \widetilde{R}_j + \delta} \approx B$*; in particular, if* $\delta \to 0$ *and* $\sum_j \widetilde{R}_j > 0$ *then* $\sum_i R_i = B$*.*
2. ***(Null-player)*** *If* $C_i(l) = 0$ *for all* l *then* $R_i = 0$*.*
3. ***(Symmetry)*** *If two clients* i, j *have identical contribution vectors* $C_i(\cdot) = C_j(\cdot)$ *(and thus identical* $|L_i| = |L_j|$*) then* $R_i = R_j$*.*
4. ***(Monotonicity / Non-decreasing)*** *Fix all* $C_k(l)$ *for* $k \neq i$*. If* $C_i(l)$ *is increased (pointwise) then* R_i *is non-decreasing; in particular* R_i *is strictly increasing except in the degenerate case where client* i *is the sole nonzero contributor and normalization pins* $R_i = B$*.*

Proof. We prove each property in turn.

Budget-balance. By construction $R_i = B \cdot \widetilde{R}_i / (\sum_j \widetilde{R}_j + \delta)$. Summing over i yields

$$\sum_{i=1}^{N} R_i \;=\; B \cdot \frac{\sum_{i=1}^{N} \widetilde{R}_i}{\sum_{j=1}^{N} \widetilde{R}_j + \delta} \;=\; B \cdot \frac{\sum_j \widetilde{R}_j}{\sum_j \widetilde{R}_j + \delta}.$$

If δ is negligible relative to $\sum_j \widetilde{R}_j$ (or set $\delta \to 0$), the right-hand side equals B, giving exact budget-balance. The formula thus enforces normalization to the available budget.

Null-player. If $C_i(l) = 0$ for all l, then $\hat{C}_i(l) = 0$ for all l, hence $S_i = 0$, hence $\widetilde{R}_i = 0$, and thus $R_i = 0$ by the normalization formula. This proves the null-player property.

Symmetry. The score S_i depends only on the vector $\{\hat{C}_i(l)\}_{l=1}^{L}$ and on $|L_i|$. If $C_i(\cdot) = C_j(\cdot)$ then $\hat{C}_i(l) = \hat{C}_j(l)$ for each l, $|L_i| = |L_j|$, so $S_i = S_j$ and therefore $\widetilde{R}_i = \widetilde{R}_j$, which implies $R_i = R_j$. Thus identical contributors receive identical rewards.

Monotonicity / Non-decreasing. Fix all other clients' contributions. For client i, each term $\hat{C}_i(l) = C_i(l)/(\bar{C}(l) + \epsilon)$ is non-decreasing in $C_i(l)$ (the denominator $\bar{C}(l)$ also increases with $C_i(l)$ but the derivative of the fraction w.r.t. $C_i(l)$ is

$$\frac{\partial}{\partial C_i(l)} \hat{C}_i(l) = \frac{(\bar{C}(l) + \epsilon) - C_i(l)}{(\bar{C}(l) + \epsilon)^2} \geq 0,$$

because $\bar{C}(l) \geq C_i(l)$. Hence $\hat{C}_i(l)$ is non-decreasing in $C_i(l)$; since $\tilde{w}(l) \geq 0$, S_i is non-decreasing in $C_i(l)$. The coverage multiplier does not depend on contribution magnitudes (only on support), so $\widetilde{R}_i = S_i \cdot \text{coverage}_i$ is non-decreasing in $C_i(l)$.

Finally, $R_i = B \cdot \widetilde{R}_i / (\sum_j \widetilde{R}_j + \delta)$. Let $T = \sum_j \widetilde{R}_j$. When $C_i(l)$ increases, $\widetilde{R}_i$ increases by some $\Delta > 0$ and T increases by the same Δ (other $\widetilde{R}_j$ are fixed). The derivative of R_i with respect to $\widetilde{R}_i$ is

$$\frac{\partial R_i}{\partial \widetilde{R}_i} = B \cdot \frac{(T + \delta) - \widetilde{R}_i}{(T + \delta)^2}.$$

□

Algorithm 2 Label Rarity–Aware Reward Computation for Federated Learning

Require: Clients C_n, labels $\mathcal{L}$, contribution scores $C_{i,l}$, budget B, rarity exponent α, small constant δ

Ensure: Rewards $R_1, \ldots, R_N$

for each label $l \in \mathcal{L}$ **do**
  $f(l) \leftarrow \left|\{\, i \mid l \in p(l)_i \,\}\right|$ {**Label Rarity Computation**}
  if $f(l) > 0$ **then**
    $w(l) \leftarrow \big(1/f(l)\big)^{\alpha}$
  else
    $w(l) \leftarrow 0$
  end if
end for
for each label $l \in \mathcal{L}$ **do**
  $C^{\max}(l) \leftarrow \max_{i:l \in L_i} C_{i,l}$ {**Per-label Shapley**}
end for
for $i = 1$ to N **do**
  $C_i^{\max} \leftarrow \sum_{l \in L_i} C^{\max}(l)$ {**Per-client Max Contributions**}
end for
for $i = 1$ to N **do**
  $\tilde{R}[i] \leftarrow 0$ {**Raw Reward Computation**}
  if $C_i^{\max} > 0$ **then**
    for each $l \in L_i$ **do**
      $\hat{C}_{i,l} \leftarrow C_{i,l} / C_i^{\max}$
      $R_i^{\text{raw}} \leftarrow \tilde{R}[i] + w(l)\, \hat{C}_{i,l}$
    end for
  end if
end for
$R_{\text{sum}} \leftarrow \sum_{i=1}^{N} \tilde{R}[i]$ {**Budget Normalization**}
for $i = 1$ to N **do**
  if $R_{\text{sum}} > 0$ **then**

$$R_i \leftarrow B \cdot \frac{\tilde{R}[i]}{R_{\text{sum}} + \delta}$$

  else
    $R_i \leftarrow B/N$
  end if
end for

return $\{R_i\}_{i=1}^{N}$