Incentivizing Black-Box Model Sharing with Fair Rewards and Payoffs

Extended Abstract

Wenyang Hu
National University of Singapore, SAP
Singapore
wenyang.hu@u.nus.edu

Xinyi Xu
National University of Singapore
Singapore
xinyi.xu@u.nus.edu

See Kiong Ng
National University of Singapore
Singapore
seekiong@nus.edu.sg

Bryan Kian Hsiang Low
National University of Singapore
Singapore
lowkh@comp.nus.edu.sg

ABSTRACT

Black-box model sharing allows multiple parties to build a high-quality ensemble model without revealing private information. However, self-interested parties require incentives, specifically Fairness and Individual Rationality, to contribute their predictions. Existing mechanisms typically handle either monetary payoffs or data rewards in isolation, failing to address scenarios where parties have varying budgets and data needs. We propose a novel incentive mechanism that fairly distributes ensemble predictions and monetary payoffs commensurate with each agent's contribution. Specifically, we use the average ensemble weight for the contribution measure and derive a closed-form solution that explicitly determines the fair reward and payoff allocation given the contribution and payment.

KEYWORDS

Collaborative Learning; Incentive Mechanisms; Black-Box Models

ACM Reference Format:
Wenyang Hu, Xinyi Xu, See Kiong Ng, and Bryan Kian Hsiang Low. 2026. Incentivizing Black-Box Model Sharing with Fair Rewards and Payoffs: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/DPVB1014

1 INTRODUCTION

Collaborative machine learning enables agents to build superior models by pooling resources. While approaches like Federated Learning [6] decentralize data, they require sharing gradients, which can leak privacy [5, 13]. *Black-box model sharing* [2, 3, 7] offers a stricter privacy guarantee: parties share only predictions $h_i(x)$ on public data, which are aggregated into ensemble prediction $h_N(x)$ as "soft labels" to improve individual models.

However, this collaboration assumes willingness to participate. In reality, agents are self-interested [9]. Without proper incentives, free-riding occurs. We identify two key incentives: **Fairness** (rewards proportional to contribution) and **Individual Rationality (IR)** (agents are better off participating than not), similar to those in incentivized data sharing [11]. Existing works struggle to balance these incentives because they treat monetary compensation and data rewards (model improvement) separately [4, 8, 11, 12]. We propose a unified 2-stage mechanism (Fig. 1): **Contribution Evaluation:** We value contributions using a *Weighted-Ensemble Game*, proving that a party's fair contribution equals its average ensemble weight; **Reward Allocation:** We introduce a *Fair-Replication Game* to jointly allocate monetary payoffs and data rewards, allowing agents to trade budget for model performance fairly.

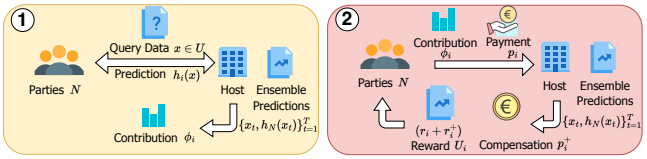

Figure 1: Overview of black-box model sharing framework. Stage 1 uses each party's prediction $h_i(x)$ to obtain its contributions ϕ_i. Stage 2 uses the contribution ϕ_i and payment p_i to realize the reward U_i (of value $r_i + r_i^+$) and compensation p_i^+.

2 METHODOLOGY

We consider n parties, each with a private model h_i trained on local data $\mathcal{D}_i$. A trusted host uses a public dataset $\mathcal{U}$ of size T to query models and aggregate predictions via an ensemble $h_N(x) = \sum_{i=1}^n \beta_{i,x} h_i(x)$, where $\beta_{i,x} \in [0,1]$ is the weight of model i for input x, determined by the selected ensemble method. The generalization error of any h on domain $\langle \mathcal{D}, f \rangle$ is $L_{\mathcal{D}}(h, f) \coloneqq \mathbb{E}_{x\sim\mathcal{D}}[\ell(h(x), f(x))]$ where ℓ is the loss function and f is the labeling function. We use the shorthand $L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h, f)$ next.

2.1 Stage 1: Valuation via Weighted Ensemble

To ensure fairness, we must quantify the marginal contribution of each party. We define the **Weighted-Ensemble Game (WEG)** where the value of a coalition $C \subseteq N$ is the total weight mass it controls: $\mathcal{V}(C) \coloneqq \frac{1}{T} \sum_{x\in U} \sum_{i\in C} \beta_{i,x}$.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/DPVB1014

Because this valuation is additive, the Shapley value [10] ϕ_i (the standard solution for fair attribution) collapses into a computationally efficient form: $\phi_i = \mathcal{V}(\{i\}) = \frac{1}{T}\sum_{x\in U}\beta_{i,x}$.

We theoretically validate this choice by analyzing the generalization error $L_{\mathcal{D}}(h_N)$. Prop. 2.1 shows that $L_{\mathcal{D}}(h_N)$ is bounded by the weighted sum of individual errors $\sum \phi_i L_{\mathcal{D}}(h_i)$. Thus, a higher ϕ_i correctly identifies a model that contributes more to reducing ensemble error.

Proposition 2.1. *For the ensemble model h_N and for any $\delta \in (0,1)$, with probability $\geq 1-n\delta$:*

$$L_{\mathcal{D}}(h_N) \leq \sum_{i=1}^{n}\left(\phi_i + \sqrt{2\log(2/\delta)/T}\right) L_{\mathcal{D}}(h_i) + \Sigma_N$$

where $\Sigma_N := \sum_{i=1}^{n}\mathbb{E}_{x\sim\mathcal{D}}[(\beta_{i,x} - \mathbb{E}_{\mathcal{D}}[\beta_{i,x}])(|h_i(x) - f(x)| - L_{\mathcal{D}}(h_i))]$.

2.2 Stage 2: Joint Allocation

Parties receive a baseline reward r_i (a subset of ensemble predictions). They may also make a monetary payment p_i to acquire additional rewards r_i^+, receiving compensation p_i^+ from the pool of others' payments. We define the updated model h' trained on a mix of private data and the received reward (ensemble predictions). We aim to satisfy:

- **Fairness:** Rewards and net payoffs are proportional to contribution ϕ_i: $\exists k_1, k_2 > 0$, s.t. $\forall i \in N, r_i = k_1\phi_i,\ r_i^+ - p_i + p_i^+ = k_2\phi_i$.
- **ϵ-Individual Rationality:** Baseline reward improves model performance, and additional reward is more valuable than its payment: $\exists \epsilon > 0$, s.t. $\forall i \in N, L_{\mathcal{D}}(h_i') - \epsilon \leq L_{\mathcal{D}}(h_i),\ r_i^+ \geq p_i$.
- **Weak Efficiency:** At least one party receives the maximum possible reward $\mathcal{V}_N$ (the full ensemble): $\exists i \in N$, s.t. $(r_i = \mathcal{V}_N)$.

To determine the fair exchange rate between money and data, we propose the **Fair-Replication Game (FRG)**, denoted $\mathcal{M}$. The value of a coalition in FRG is defined by $\mathcal{M}(C) = (\sum_{i\in C}\phi_i)\times\gamma$, where $\gamma = \sum_{j\in N} p_j/(\mathcal{V}_N - \phi_j)$ represents the total monetary pool normalized by contribution scarcity.

We analyze the combined game $\mathcal{V}+\mathcal{M}$. We derive the final utility u_i (the combined Shapley value), which specifies the allocation:

Theorem 2.2. *Given $\mathcal{V}$ and $\mathcal{M}$, the utility $u_i := \phi_i(\mathcal{V}+\mathcal{M})$ for each party $i \in N$ can be decomposed as $u_i = r_i + r_i^+ + p_i^+ - p_i$ where $r_i = \phi_i$ is the Shapley value in game G, specifically as follows,*

$$u_i = r_i + \underbrace{\frac{\mathcal{V}_N \times p_i}{\mathcal{V}_N - \phi_i}}_{r_i^+} + \underbrace{\sum_{j\in N\setminus\{i\}}\frac{\phi_i \times p_j}{\mathcal{V}_N - \phi_j}}_{p_i^+} - p_i.$$

The utility u_i satisfies (a) payoff balance: $\sum_{i\in N}(p_i^+ - p_i) = 0$, (b) dummy payment: $\forall C \subseteq N\setminus\{i\},\ \mathcal{V}(C\cup i) = \mathcal{V}(C) \Rightarrow u_i = 0$, $r_i^+ = p_i$, (c) semi-symmetry: $\forall C \subseteq N\setminus\{i,j\},\ \mathcal{V}(C\cup i) = \mathcal{V}(C\cup j) \Rightarrow u_i = u_j$, and (d) strict monotonicity: $(\exists j \in N, p_j' > p_j) \wedge (\forall k \in N\ p_k' \geq p_k) \Rightarrow \forall i \in N, u_i' > u_i$.

Realization Scheme. Based on Thm. 2.2, we now describe how the host allocates rewards and payoffs in practice. Party i receives a total reward of value $r_i + r_i^+ = r_i + \mathcal{V}_N \times p_i/(\mathcal{V}_N - \phi_i)$, and a net payoff of value $p_i^+ - p_i = \sum_{j\in N\setminus\{i\}}\frac{\phi_i\times p_j}{\mathcal{V}_N - \phi_j} - p_i$ from the host. The host samples a subset $\mathcal{S}_i \subseteq \mathcal{U}$ with $|\mathcal{S}_i| = T_i = (r_i + r_i^+)\times T$ and send reward $\mathcal{R}_i = \{(x, h_N(x))\}_{x\in\mathcal{S}_i}$ and payoff p_i^+ to party i. Parties fine-tune h_i on $(1-\alpha_i)\mathcal{D}_i \cup \alpha_i\mathcal{R}_i$ to produce h_i'.

2.3 Incentive Guarantee

The allocation based on Thm. 2.2 guarantees our desired incentives: Fairness and Weak Efficiency, by using the scaled reward $r_i = (\phi_i/\phi^*)\times\mathcal{V}_N$. Additionally, $r_i^+ \geq p_i$ is always satisfied.

We now elaborate $L_{\mathcal{D}}(h_i') - \epsilon \leq L_{\mathcal{D}}(h_i)$ by analyzing ϵ which is a virtual regret that needs to be minimized to incentivize participation. We first define the empirical risk trained on the data mix $\hat{L}_{\mathcal{D}_i'}(h) := (1-\alpha_i)\hat{L}_{\mathcal{D}_i}(h) + \alpha_i\hat{L}_{\mathcal{D}}(h, h_N)$ and $\alpha_i \in [0,1]$ the mixing weight for balancing training data. By leveraging technical results from domain learning theory [1] and utilizing the *ensemble domain* $\langle\mathcal{D}, h_N\rangle$, we provide a specific result of ϵ-IR in Prop. 2.3.

Proposition 2.3. *Let $\mathcal{H}$ be a hypothesis space of VC dimension d . Given m_i the data size of $\mathcal{D}_i$, and a distribution divergence measure $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}) := 2\sup_{h,h'\in\mathcal{H}}\left|L_{\mathcal{D}_i}(h,h') - L_{\mathcal{D}}(h,h')\right|$, with probability at least $1-n\delta$, ϵ-IR is satisfied such that $\forall i \in N\ L_{\mathcal{D}}(h_i') - \epsilon \leq L_{\mathcal{D}}(h_i)$ with $\epsilon = \max_{i\in N}\epsilon_i$ and*

$$\epsilon_i = A\left(\alpha_i^2/T_i + (1-\alpha_i)^2/m_i\right)^{\frac{1}{2}} + \alpha B + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D})$$

where $A = 4\sqrt{2d\log(2(T_i+m_i+1)) + 2\log(\frac{8}{\delta})}$ *and* $B = -d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}) + 2L_{\mathcal{D}}(h_N)$.

As $\epsilon = \max_{i\in N}\epsilon_i$, minimizing ϵ is equal to minimize each ϵ_i. Thus, a smaller ϵ_i for all $i \in N$ indicates a stronger ϵ-IR guarantee. Prop. 2.3 shows that ϵ_i mainly depends on the ensemble error $L_{\mathcal{D}}(h_N)$, the mixing value α_i, and its reward size T_i, as $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D})$ is fixed. For a given collaboration, $L_{\mathcal{D}}(h_N)$ (due to ensemble method) and T_i (due to fixed ϕ_i and p_i) are also fixed. When $\alpha_i = 1$, it implies h_i' is only trained on the ensemble predictions, and $\epsilon_i(1)$ mainly depends on $L_{\mathcal{D}}(h_N)$ the error of ensemble predictions; when $\alpha_i = 0$, h_i' is the same as h_i, and $\epsilon_i(0)$ mainly depends on $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D})$ the distribution divergence. We can write $\epsilon_i(\alpha_i)$ as a function of α_i, which achieves its minimum by using the optimal value α^*. We then have the strongest ϵ-IR guarantee where $\epsilon = \max_{i\in N}\epsilon_i$ is minimized. In the ideal case if $\epsilon_i = 0\ \forall i \in N$, the strict IR (i.e., $L_{\mathcal{D}}(h_i') \leq L_{\mathcal{D}}(h_i)$) is satisfied; this holds only when each party has infinite optimal ensemble predictions (i.e., $T_i = \infty$ and $L_{\mathcal{D}}(h_N) = 0$) or infinite optimal source data (i.e., $m_i = \infty$ and $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i, \mathcal{D}) = 0$), which is hard to satisfy in practice. In practice, the virtual regret ϵ is not needed, and the strict IR is consistently achieved empirically, underscoring that parties are not made worse off by the collaboration. The α_i^* not only results in the strict IR, but also yields the largest performance improvement empirically.

3 CONCLUSION

In this work, we presented a theoretically grounded mechanism for black-box model sharing, with a close-form solution to jointly allocate rewards and monetary payoffs, satisfying Shapley fairness and guarantees ϵ-individual rationality.

ACKNOWLEDGMENTS

This research/project is supported by SAP and Singapore's Economic Development Board under the Industrial Postgraduate Programme.

REFERENCES

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1 (2010), 151–175.

[2] Hongyan Chang, Virat Shejwalkar, Reza Shokri, and Amir Houmansadr. 2021. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. In *Proc. of the 1st NeurIPS Workshop on New Frontiers in Federated Learning*.

[3] Haozhe Feng, Zhaoyang You, Minghao Chen, Tianye Zhang, Minfeng Zhu, Fei Wu, Chao Wu, and Wei Chen. 2021. KD3A: Unsupervised Multi-Source Decentralized Domain Adaptation via Knowledge Distillation. In *Proc. ICML*, Vol. 139. 3274–3283.

[4] Dongge Han, Michael Wooldridge, Alex Rogers, Olga Ohrimenko, and Sebastian Tschiatschek. 2023. Replication Robust Payoff Allocation in Submodular Cooperative Games. *IEEE Transactions on Artificial Intelligence* 4, 5 (2023), 1114–1128.

[5] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. 2017. Deep models under the GAN: information leakage from collaborative deep learning. In *Proc. SIGSAC (CCS '17)*. 603–618.

[6] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* 14, 1–2 (2021), 1–210.

[7] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. 2020. Ensemble distillation for robust model fusion in federated learning. In *Proc. NeurIPS*, Vol. 33. 2351–2363.

[8] Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. 2021. Dealer: an end-to-end model marketplace with differential privacy. In *Proc. VLDB Endow.*, Vol. 14. 957–969.

[9] Aniket Murhekar, Zhuowen Yuan, Bhaskar Ray Chaudhury, Bo Li, and Ruta Mehta. 2023. Incentives in Federated Learning: Equilibria, Dynamics, and Mechanisms for Welfare Maximization. In *Proc. NeurIPS*, Vol. 36. 17811–17831.

[10] L. S. Shapley. 1953. A value for n-person games. In *Contributions to the Theory of Games*, H. W. Kuhn and A. W. Tucker (Eds.). Vol. 2. Princeton University Press, 307–317.

[11] Rachael Hwee Ling Sim, Yehong Zhang, Mun Choon Chan, and Bryan Kian Hsiang Low. 2020. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*. 8927–8936.

[12] Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo, and Bryan Kian Hsiang Low. 2022. Incentivizing collaboration in machine learning via synthetic data rewards. In *Proc. AAAI*, Vol. 36. 9448–9456.

[13] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Proc. NeurIPS*, Vol. 32.