Poster: Decentralized Trustless Query-Based Collaborative Learning under Adversarial Heterogeneity without Model Sharing

Ali Mohammadi Ruzbahani
Smart Cyber-Physical (SCPS) Lab
University of Calgary
Email: ali.mohammadiruzbaha@ucalgary.ca

Abbas Yazdinejad
DCAIlab
University of Regina
Email: Abbas.Yazdinejad@uregina.ca

Hadis Karimipour
Smart Cyber-Physical (SCPS) Lab
University of Calgary
Email: hadis.karimipour@ucalgary.ca

***Abstract*—Federated learning (FL) is poorly suited to adversarial multi-party settings because it depends on trusted aggregation and gradient-based interaction, both of which increase exposure to poisoning and information leakage. We present a decentralized collaborative learning framework that replaces parameter exchange with privacy-preserving query-based interaction, enabling heterogeneous participants to improve jointly without exposing raw data, gradients, or model parameters. Participants issue queries derived from local failure cases and receive aggregated predictive feedback through Byzantine-resilient consensus. Preliminary results on 15 heterogeneous nodes under severe non-IID conditions show substantial gains over local-only training, improving mean accuracy by +0.2036 and worst-node performance by +0.3032. These findings suggest that query-based interaction is a promising alternative to parameter sharing in adversarial settings.**

1. Introduction

Collaborative machine learning is important in security-critical settings such as cross-organizational threat intelligence sharing, collaborative intrusion detection, and joint fraud analytics. In these scenarios, participants can benefit from one another's predictive signals, yet cannot trust a central coordinator with training data from other parties and are unwilling to share model parameters, gradients, or raw data.

Current federated learning (FL) solutions are ill-suited to this threat model. A centralized aggregation point creates a single point of failure and introduces attack opportunities such as poisoning and backdoors, while gradient-based interaction can leak sensitive information through reconstruction attacks [1]. Many existing approaches are also poorly matched to these settings because they assume architectural homogeneity, shared public datasets, or coordination models that are difficult to sustain across independent and potentially competing organizations. FedAvg [2] relies on centralized aggregation; DS-FL and FedMD require shared public data and coordinated interaction [3], [4]; and Krum improves Byzantine robustness only under trusted centralized coordination [5]. This leaves open the question: ***how can mutually distrustful and potentially malicious parties collaboratively improve their models without exposing internal states, relying on centralized coordination, or assuming architectural compatibility?***

In this work, we address this question by rethinking collaborative learning through query-based interaction rather than parameter sharing. This enables decentralized collaboration across heterogeneous models while reducing communication overhead and improving protocol-level privacy and accountability.

Contributions: We present a trustless collaborative learning framework that replaces gradient exchange with a query-based protocol for secure knowledge sharing through privacy-preserving Q/A and Byzantine-resilient consensus. The framework provides trustless collaboration without model sharing, reducing attack surfaces and eliminating gradient leakage; and heterogeneity-aware interaction that enables secure collaboration across dissimilar model architectures without shared data.

2. Framework and Security Properties

Threat Model: We consider N nodes owned by mutually distrustful organizations. *Participant nodes* are honest-but-curious: they follow the protocol correctly but may attempt to infer private information from received queries. *Committee members* may be fully Byzantine: up to $f < N/3$ may return arbitrary responses, collude, or selectively drop messages. The adversary observes all protocol traffic but cannot break the privacy of the Laplace mechanism. No node shares raw data, gradients, or model weights at any point. Fig. 1 summarizes the overall workflow.

privacy-preserving query exchange: When node i misclassifies a local sample $x \in D_i$, it constructs a privacy-preserving query from anchor-relative embedding distances: $\mathbf{q}_i = \left[d(z_i, a_c)\right]_{c=1}^{C} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathrm{Lap}\left(\frac{\Delta_f}{\varepsilon}\right)^C$, where $z_i = \mathrm{enc}_i(x)$ is the local embedding of x. Here, $\{a_c\}_{c=1}^{C}$ are defined as fixed, class-wise orthonormal basis vectors in a common hyperspace, providing a universal reference frame without requiring shared data. To ensure semantic consistency across heterogeneous architectures, each node independently aligns its local latent space to these anchors during training via an auxiliary projection loss. This mech-

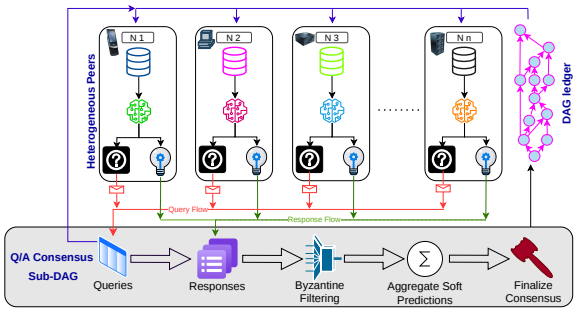


Figure 1. Proposed trustless collaborative learning framework.

TABLE 1. CIFAR-10 RESULTS AT ROUND 70.

Configuration	Mean Acc.	Std	Worst Node
Local-only training	0.2865	0.0853	0.1000
Q/A collaborative	0.4901	0.0424	0.4032
Change	+0.2036	-0.0429	+0.3032

anism satisfies ε-local differential privacy under standard Laplace perturbation assumptions [6], masking the exact relative position of the sample. Peer nodes interpret $\mathbf{q}_i$ by identifying regions in their own embedding spaces that exhibit similar relative distances to the public anchors, returning soft class distributions. A rotating Byzantine-resilient committee filters these responses and records a finalized consensus on the DAG ledger. The querying node then incorporates this feedback into local retraining, achieving knowledge transfer without exposing parameters or raw data.

Robustness and accountability: The framework uses a protocol-level committee to filter and finalize query responses before recording outcomes on the DAG ledger. Under the assumption that fewer than one-third of committee members are adversarial, committee-based aggregation provides Byzantine resilience, while Sybil behavior is discouraged through bonded identities and capped committee influence.

3. Preliminary Results

The framework is evaluated on CIFAR-10 with 15 heterogeneous nodes under high data skewness ($\alpha = 0.1$). Incorporating a diverse architectural suite (ResNet, DenseNet, MobileNet, ShuffleNet, GoogLeNet, and VGG), we compare the Q/A collaborative protocol against a local-only baseline. Direct benchmarking against parameter-sharing methods like FedAvg [2] is fundamentally incompatible with our architectural heterogeneity and privacy-preserving constraints. Consequently, more complex baselines, including Knowledge Distillation-based ensembles, are reserved for comprehensive full-paper analysis.

Table 1 summarizes the final-round results. At round 70, the proposed framework improves mean accuracy by $+0.2036$ and worst-node performance by $+0.3032$ over local-only training, while reducing the standard deviation

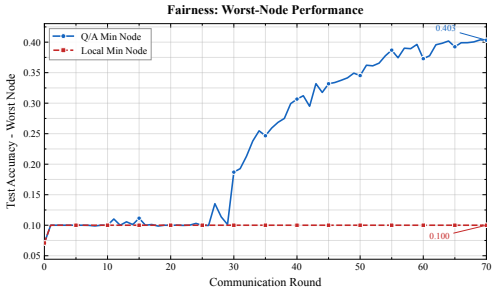


Figure 2. Worst-node test accuracy over training rounds. Q/A collaboration substantially improves the weakest participant compared to local-only training.

across nodes from 0.0853 to 0.0424. Fig. 2 further shows that these gains are not limited to average-case performance: the weakest node improves steadily during training under Q/A collaboration, whereas the worst node in the local-only setting remains near chance level throughout.

This behavior indicates that the protocol distributes benefit broadly across heterogeneous participants rather than selectively favoring particular architectures. The effect is also stable late in training: Q/A collaboration outperforms the local-only baseline in all 10 of the final 10 rounds. Overall, these preliminary results suggest that compact privacy-preserving query exchange can provide effective knowledge transfer under severe heterogeneity without gradient or parameter sharing.

4. Conclusion

These results demonstrate that privacy-preserving query-based interaction provides a robust alternative to parameter-sharing approaches in adversarial settings.

References

[1] L. Zhu, Z. Liu, and S. Han, "Deep Leakage from Gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 14747–14756.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 54, 2017, pp. 1273–1282.

[3] S. Itahara, T. Nishio, Y. Koda, M. Morikura, and K. Yamamoto, "Distillation-Based Semi-Supervised Federated Learning for Communication-Efficient Collaborative Training With Non-IID Private Data," in *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 191–205, Jan. 2023, doi: 10.1109/TMC.2021.3070013.

[4] D. Li and J. Wang, "FedMD: Heterogenous Federated Learning via Model Distillation," arXiv:1910.03581, 2019.

[5] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, pp. 119–129.

[6] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local Privacy and Statistical Minimax Rates," in *Proc. 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013, pp. 429–438.

Decentralized Trustless Query-Based Collaborative Learning under Adversarial Heterogeneity without Model Sharing

Ali Mohammadi Ruzbahani [1] Abbas Yazdinejad [2] Hadis Karimipour [1]

[1]Smart Cyber-Physical (SCPS) Lab, University of Calgary, {ali.mohammadiruzbaha, hadis.karimipour}@ucalgary.ca [2]DCAllab, University of Regina, Abbas.Yazdinejad@uregina.ca

Motivation and Problem

Collaborative learning is increasingly needed in security-critical settings, but standard federated learning is poorly matched to adversarial multi-party environments due to centralized coordination, privacy leakage through gradient exchange, and restrictive assumptions such as architectural homogeneity or shared public data.

Comparison with Prior Work

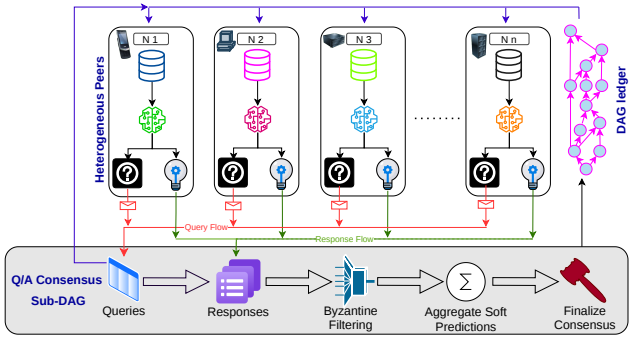
Property	FedAvg	DS-FL	FedMD	Krum	Ours
Decentralized	✗	✗	✗	✗	✓
No Gradient Sharing	✗	✓	✓	✗	✓
No Shared Public Data	✓	✗	✗	✓	✓
Model Heterogeneity	✗	✓	✓	✗	✓
Byzantine-Resilient	✗	✗	✗	✓	✓
Formal Privacy (DP)	✗	✗	✗	✗	✓
Auditability via Ledger	✗	✗	✗	✗	✓

Research Question

How can distrustful and adversarial parties collaborate effectively without sharing internal states, relying on centralized coordination, or assuming architectural homogeneity?

DTQCL Framework

N nodes collaborate through a rotating committee, with fewer than one-third of committee members assumed Byzantine. Nodes never share raw data, gradients, or model weights, and finalized outcomes are recorded on a publicly verifiable DAG ledger.



Core Mechanisms

Privacy-preserving query release. For fixed orthonormal anchors $\{a_c\}_{c=1}^{C} \subset \mathbb{R}^d$ and $z_i = \mathrm{enc}_i(x)$ with $\|z_i\|_2 \le R$,

$$q_i = \phi(z_i) + \eta_i, \qquad \phi(z_i) = \big[d(z_i, a_c)\big]_{c=1}^{C}, \qquad \eta_i \sim \mathrm{Lap}\Big(\frac{\Delta_\phi}{\varepsilon}\Big)^{C},$$

where Δ_ϕ is the global ℓ_1-sensitivity of ϕ. This yields ε-local differential privacy.

Cross-architecture anchor alignment.

$$\mathcal{L}_i(x, y) = \mathcal{L}_{\mathrm{task}}(f_i(x), y) + \lambda_{\mathrm{align}} \|g_i(h_i(x)) - a_y\|_2^2,$$

with backbone h_i, projection head g_i, and class anchor a_y. This induces a shared geometric reference frame across heterogeneous architectures.

Robust committee aggregation and finalization.

$$\hat{p}_i = \mathrm{TrimMean}\Big(\{p_j(q_i)\}_{j \in \mathcal{R}_i^{(t)}}\Big), \qquad |\mathcal{F}_t| < \tfrac{1}{3}|\mathcal{C}_t|,$$

where $\mathcal{F}_t \subseteq \mathcal{C}_t$ denotes Byzantine committee members.

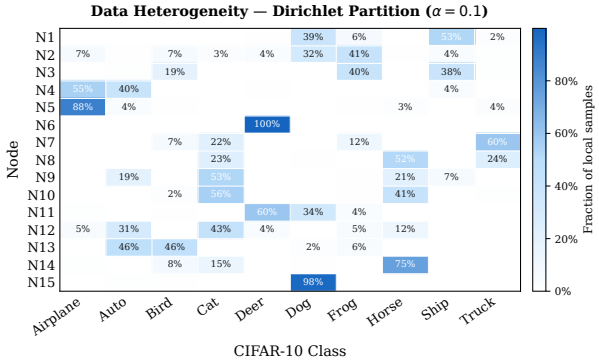
Ledger-backed auditability.

$$(i,\ H(q_i),\ H(\hat{p}_i),\ t)$$

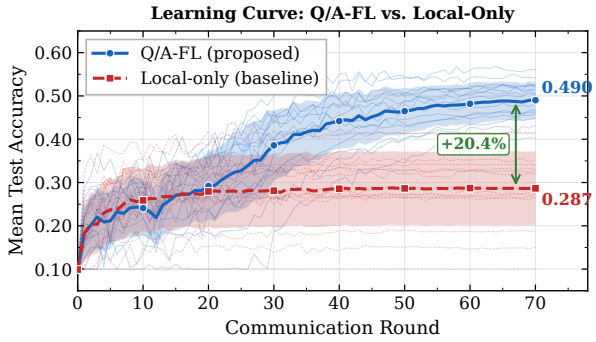
is appended to the DAG ledger for public auditability.

Experimental Setup

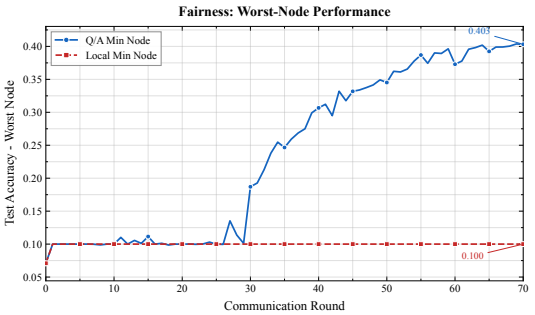
DTQCL is evaluated on **CIFAR-10** using **15 heterogeneous nodes** under **severe Dirichlet non-IID partitioning** ($\alpha = 0.1$) for **70 rounds**; **Q/A collaboration** is compared against **local-only training**.



Preliminary Results



Config.	Mean	Std	Worst
Local-only	0.2865	0.0853	0.1000
Q/A collab.	0.4901	0.0424	0.4032
Change	+0.2036	-0.0429	+0.3032



Conclusion

DTQCL enables trustless, decentralized collaborative learning without model sharing. Through privacy-preserving query/answer interaction, anchor-based alignment for heterogeneous models, and Byzantine-resilient committee aggregation with DAG-backed auditability, it achieves effective and robust knowledge transfer under adversarial heterogeneity. Experiments under severe non-IID settings show consistent gains in average accuracy, weakest-node performance, and stability.