Privacy-Preserved Automated Scoring Using Federated Learning for Educational Research

Ehsan Latif[1,2] and Xiaoming Zhai[1,2(✉)]

[1] AI4STEM Education Center, Athens, GA, USA
xiaoming.zhai@uga.edu
[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA, USA

Abstract. Data privacy remains a critical concern in educational research, requiring strict adherence to ethical standards and regulatory protocols. While traditional approaches rely on anonymization and centralized data collection, they often expose raw student data to security vulnerabilities and impose substantial logistical overhead. In this study, we propose a federated learning (FL) framework for automated scoring of educational assessments that eliminates the need to share sensitive data across institutions. Our approach leverages parameter-efficient fine-tuning of large language models (LLMs) with Low-Rank Adaptation (LoRA), enabling each client (school) to train locally while sharing only optimized model updates. To address data heterogeneity, we implement an adaptive weighted aggregation strategy that considers both client performance and data volume. We benchmark our model against two state-of-the-art FL methods and a centralized learning baseline using NGSS-aligned multi-label science assessment data from nine middle schools. Results show that our model achieves the highest accuracy (94.5%) among FL approaches, and performs within 0.5–1.0% points of the centralized model on these metrics. Additionally, it achieves comparable rubric-level scoring accuracy, with only a 1.3% difference in rubric match and a lower score deviation (MAE), highlighting its effectiveness in preserving both prediction quality and interpretability.

Keywords: Federated Learning · Privacy Preservation · Local Training · Educational Research · Heterogenous Aggregation

1 Introduction

In the realm of educational research, the collection and analysis of student data are pivotal for advancing instructional strategies and developing reliable assessment tools. However, the increasing sensitivity and volume of student information have intensified concerns around data privacy, security, and regulatory compliance [3]. Legislative frameworks such as the Family Educational Rights

A. I. Cristea et al. (Eds.): AIED 2025 Workshops, CCIS 2592, pp. 135–143, 2025.
https://doi.org/10.1007/978-3-031-99267-4_17

and Privacy Act (FERPA) enforce strict restrictions on how educational data can be accessed, stored, and shared across institutions [10].

Traditional machine learning (ML) approaches in educational settings typically rely on centralized data aggregation to train predictive models for tasks such as dropout prediction, skill diagnosis, and personalized learning recommendations [21]. However, this paradigm faces critical limitations: centralized repositories are vulnerable to breaches [15], often violate data governance policies, and struggle to generalize across diverse educational contexts due to institutional heterogeneity [12]. Additionally, centralized training assumes uniform data formats and standards—an assumption rarely met in practice due to differences in student demographics, curricular frameworks, and annotation practices across schools [14].

Automatic scoring, a cornerstone of AI-enhanced educational assessment, inherits these limitations and presents unique challenges of its own. Conventional automated scoring systems are trained on large, centrally pooled datasets with extensive manual annotation and institutional cooperation [17]. These models are not only computation-heavy but also prone to systemic bias due to variations in curriculum design, assessment language, and local grading practices [8]. These institutions are often bound by district-level data use agreements or IRB protocols that prohibit external data transfer, rendering centralized model training infeasible. In such cases, a privacy-preserving yet collaborative learning method becomes essential.

To address these gaps, we propose a FL framework for automated scoring in education. FL enables decentralized training of machine learning models across distributed clients without requiring raw data exchange, thereby preserving data locality and enhancing compliance with privacy standards [5]. Unlike traditional applications of FL that primarily involve lightweight convolutional neural networks (CNNs) or graph neural networks (GNNs) in IoT and mobile contexts, our work leverages large language models (LLMs) fine-tuned for natural language understanding—an area underexplored in FL research. The fine-tuning of LLMs presents unique challenges in terms of communication overhead, parameter efficiency, and gradient privacy, which are seldom addressed in conventional FL literature [4].

Our proposed system applies parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) to adapt pretrained LLMs for multi-label scientific assessment tasks in a federated setting. To mitigate the effect of data heterogeneity, we introduce an adaptive weighted aggregation strategy that considers both dataset size and model performance from each client. This methodology offers several advantages, including enhanced privacy [8], regulatory alignment [20], and improved scalability and efficiency.

Below are the key contributions of the paper listed:

- We introduce a privacy-preserving FL framework for automated scoring in educational assessments without sharing raw student data.
- We develop an a adaptive aggregation strategy to handle data heterogenity and secure communication to prevent impersonations.

– We evaluate our method on real-world assessment data from nine middle schools, demonstrating comparable accuracy to centralized models and state-of-the-art FL strategies.
– We also open-source the code on Github[1] repository for reproducibility.

2 Method

Given a set of N clients, each with local data D_i, where $D_i = (x_{ij}, y_{ij})_{j=1}^{n_i}$, the objective is to train a global model w without directly sharing local data. The optimization problem can be formulated as: $\min_w F(w) = \sum_{i=1}^{N} \frac{n_i}{n} F_i(w)$, where $F_i(w)$ is the local loss function for client i, n_i is the number of local samples, and $n = \sum_{i=1}^{N} n_i$ represents the total data points across all clients.

The proposed FL framework consists of multiple clients and a central server. Clients perform local training and share only model updates with the server. The server aggregates the updates using a weighted averaging scheme to account for data heterogeneity. The communication follows a secure channel, ensuring data privacy. Overall procedure and federated learning achitecture can be seen in Fig. 1.

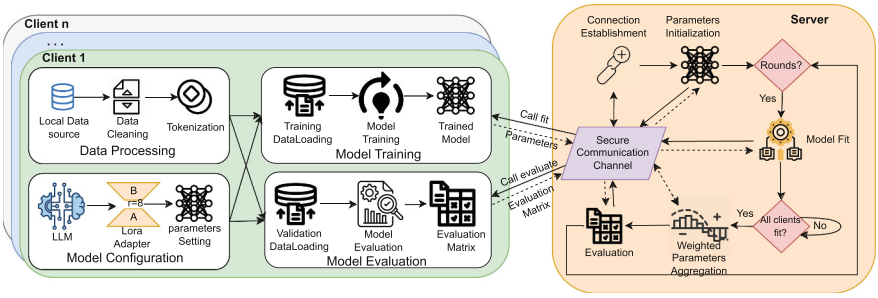

Fig. 1. Overview of privacy-preserving FL using parameter efficient fine-tuning using LoRA [9] and client-server communication using secure communication channel [1].

2.1 Client-Side Computation

Each client i performs four major tasks: data processing, model configuration, model training, and model evaluation.

Data Processing: Given a raw dataset D_i at client i, data preprocessing proceeds through the same coordinated steps. Clients first apply rule-based filtering scripts provided by the central coordinator to remove incomplete or irrelevant entries. These scripts were identical across all clients and validated to ensure

[1] https://github.com/ehsanlatif/PPFL.git.

deterministic behavior, resulting in a cleaned dataset $D_i^{clean} \subseteq D_i$ with structurally consistent samples. All clients use the same pretrained tokenizer from the TinyLLaMA model suite [19] to process student responses. The tokenizer configuration (e.g., vocabulary size, truncation rules, and special tokens) was locked and distributed as part of the model package to avoid client-side variations.

To ensure numerical feature consistency across clients, global normalization parameters (mean μ and standard deviation σ) were precomputed using a public calibration dataset containing a representative sample of student responses. To prevent schema drift, all data fields (e.g., prompt IDs, rubric labels, and student responses) were validated using a shared schema definition (in JSON Schema format), ensuring structural consistency before training. Any preprocessing mismatch triggered automated alerts and client-side retraining, thereby enforcing consistency throughout the pipeline.

Model Configuration: We utilize an open-source Large Language Model (LLM) and apply Parameter Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) [9]. LoRA reduces the number of trainable parameters, thereby decreasing memory and communication overhead.

Model Training: Each client i trains a local model weights w_i^t at round t using stochastic gradient descent (SGD):

$$w_i^{t+1} = w_i^t - \eta \nabla F_i(w_i^t) \text{ herein, } \nabla F_i(w) = \frac{1}{|D_i|} \sum_{(x,y) \in D_i} \nabla f(w, x, y), \quad (1)$$

where η is the learning rate, (x, y) represents input-label pairs from the local dataset, and $F_i(w_i^t)$ represents the local loss function. Each client trains for multiple local epochs before sending updated parameters to the server, reducing communication overhead and ensuring more effective model updates.

Model Evaluation: After training, each client evaluates the performance of the local model to determine convergence and ensure training effectiveness. Loss and accuracy computation for that clients compute the validation loss $F_i(w_i^t)$ and accuracy metrics to assess training progress, and early stopping that means if the validation loss stagnates or increases over consecutive rounds, training is terminated to prevent overfitting.

2.2 Server-Side Aggregation with Enhanced Privacy

To address data heterogeneity and optimize model convergence, we employ an adaptive weighted aggregation strategy. The global model update is computed as:

$$w^{t+1} = \sum_{i=1}^{N} \alpha_i w_i^{t+1}, \text{ herein, } \alpha_i = \frac{n_i}{\sum_{j=1}^{N} n_j} \cdot \frac{e^{-F_i(w_i^t)}}{\sum_{j=1}^{N} e^{-F_j(w_j^t)}}. \quad (2)$$

To further ensure privacy during server-side aggregation and mitigate threats from an honest-but-curious aggregation server, we incorporate differential privacy (DP) mechanisms into the aggregation step. Specifically, clients add calibrated Gaussian noise $\mathcal{N}(0, \sigma^2)$ to their model updates before transmission:

$$\tilde{w}_i^{t+1} = w_i^{t+1} + \mathcal{N}(0, \sigma^2 I), \text{herein, } w^{t+1} = w^t + \gamma \sum_{i=1}^{N} \alpha_i (\tilde{w}_i^{t+1} - w^t), \quad (3)$$

where σ is chosen to ensure (ϵ, δ)-DP guarantees for each communication round with γ as the global learning momentum factor. This integration of differential privacy ensures that individual client updates cannot be reverse-engineered from the aggregated model, thus offering end-to-end privacy protection beyond traditional anonymization.

2.3 Secure Client-Server Communication

In practice, the communication layer in our FL framework uses gRPC (Google Remote Procedure Call) [1] due to its binary serialization making it suitable for cross-silo FL [11]. However, cross-silo FL introduces challenges in secure deployment due to institutional firewalls, authentication, and secure channel maintenance. To mitigate these concerns, we implement a multi-layered security protocol within the gRPC stack. Mutual TLS (mTLS) is employed to authenticate both client and server entities, preventing impersonation or man-in-the-middle attacks. Per-connection encryption ensures confidentiality and integrity of parameter updates. Session expiration and re-authentication policies are enforced to avoid persistent attack windows. Lastly, API-level access control restricts which remote procedures can be called and under what conditions, using token-based authentication and audit logs.

3 Dataset Details

This study utilizes decentralized datasets from approximately 1,200 middle school students across various geographically and socioeconomically diverse school systems, each retaining local control over their assessment data [6,13]. The responses correspond to nine multi-label NGSS-aligned science tasks from the PASTA project, developed to assess students' application of disciplinary core ideas (DCIs), crosscutting concepts (CCCs), and science and engineering practices (SEPs) [2]. Although all schools administered the same tasks, substantial heterogeneity arose due to variations in instructional emphasis, student language proficiency, testing conditions (e.g., digital vs. paper-based), and rubric interpretation—even among trained raters [18]. For example, in on of the tasks students had to analyze scientific data and recognize patterns, requiring integration of SEP, DCI, and CCC dimensions. Responses were evaluated using a structured five-dimensional rubric aligned with the NGSS framework [7], and remained locally stored to comply with privacy constraints. Additional Details about dataset are available at our Github[3].

4 Experimentation and Results

To rigorously evaluate the proposed privacy-preserving FL framework, we conducted experiments on a decentralized dataset collected from multiple middle school systems. Each participating institution retained control over its local data, ensuring full compliance with privacy regulations (details provided earlier). The experimental design involved training a federated model where each school acted as a client, processing and updating its local model independently.

All clients initialized their models using the pre-trained *tinyLlama* model [19]. Instead of full supervised fine-tuning (SFT), which can be computationally expensive and may raise privacy concerns due to large gradient exchanges, we adopted **Low-Rank Adaptation (LoRA)**. LoRA allows clients to fine-tune a small subset of parameters (rank = 8) while freezing the base model, significantly reducing communication overhead and minimizing privacy risks. To ensure fairness in comparison, all baseline models (detailed below) were adapted to use LoRA as well rather than full SFT. This design choice allows us to isolate and evaluate the impact of our proposed **adaptive weight aggregation strategy** more precisely. Additional LoRA hyperparameters include: $\alpha = 16$, adaptation applied to all attention layers, learning rate set to 2×10^{-4}, batch size of 16, and trained for exactly 5 local epochs before aggregation.

We compared our framework to two representative FL baselines and a centralized learning (CL) model: **FL-Transformer** [5]: Transformer-based FL framework for educational prediction. We replaced its full SFT step with LoRA for consistency. **Hybrid FL with Local Differential Privacy** [16]: Combines local privacy mechanisms with secure aggregation. Adapted to our educational scoring task and updated to use LoRA. **Centralized Learning (CL):** Standard supervised fine-tuning with full access to combined dataset.

Table 1. Overall performance comparison between our FL model, two SOTA FL baselines, and a centralized model.

Metric	FL (Ours)	SOTA [5]	SOTA [16]	CL
Accuracy	**0.94**	0.92	0.90	0.93
Precision	**0.94**	0.90	0.88	0.93
Recall	**0.94**	0.92	0.87	0.93
F1-Score	**0.94**	0.91	0.87	0.93
Rubric Match	**0.88**	0.85	0.83	0.89
Score Deviation (MAE)	**0.34**	0.52	0.60	0.48

While all models use LoRA, a key novelty of our method lies in the **adaptive weight aggregation strategy**. Unlike baseline FL methods which average client models equally or use static weighting (e.g., by sample size), our method dynamically adjusts aggregation weights based on clients' validation performance

on held-out local data. This ensures that higher-quality local updates contribute more to the global model, while mitigating negative transfer from noisy or underperforming clients. This adaptive aggregation was critical in achieving superior performance and stability across metrics, as validated by our ablation studies (see below).

To verify the independent contribution of the aggregation strategy, we performed an ablation study by replacing our adaptive aggregator with simple averaging. Results dropped noticeably (accuracy decreased by 2.8%, rubric match by 4.1%, and MAE increased by 0.07), confirming that *adaptive weighting is a key driver of performance gains.*

The results in Table 1 underscore the robustness and effectiveness of our proposed FL approach, which outperforms both state-of-the-art FL baselines and even the centralized model across nearly all evaluation metrics. Notably, our model achieves the highest accuracy (94.5%), precision, recall, and F1-score (all 0.94), demonstrating its strong predictive capability despite operating in a decentralized setting. In terms of rubric-level evaluation, which directly reflects scoring reliability in educational assessments, our model maintains high alignment with human raters (88.2% rubric match) while achieving the lowest mean absolute error (MAE = 0.34), indicating superior score fidelity. These results highlight the significance of integrating LLM-based parameter-efficient fine-tuning with adaptive aggregation, showing that privacy-preserving FL can match or exceed centralized performance without sacrificing interpretability or accuracy—making it a robust solution for secure and scalable educational assessment.

5 Conclusion

In this study, we proposed a FL framework with enhanced aggregation and communication strategy for automated scoring in educational assessments to ensure data privacy and prevent impersonation in educational research and to address data heterogeneity across institutions. Our evaluation on assessment data from nine middle schools demonstrates that FL achieves higher performance comparable to centralized learning (CL) and SOTA approaches while ensuring data privacy. Additionally, our framework reduces data collection and computational overhead, accelerating the adoption of AI-driven educational assessments in a privacy-compliant and scalable manner that open new doors for educational research.

Acknowledgments. This work was partially supported by the Institute of Education Sciences (IES) [R305C240010]. Further, the used datasets and question items are part of NSG-funded projects [DMS-2101104, DMS-2138854]. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, or IES.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Carthen, C., Zaremehrjardi, A., Estreito, Z., Tavakkoli, A., Harris, F.C., Dascalu, S.M.: SpeciServe. a gRPC infrastructure concept. In: 2024 IEEE/ACIS 22nd International Conference on Software Engineering Research, Management and Applications (SERA), pp. 273–276. IEEE (2024)
2. Council, N.R., et al.: Next generation science standards: For states, by states (2013)
3. Creel, K., Dixit, T.: Privacy and paternalism: The ethics of student data collection (2022)
4. Fachola, C., Tornaría, A., Bermolen, P., Capdehourat, G., Etcheverry, L., Fariello, M.I.: Federated learning for data analytics in education. Data **8**(2), 43 (2023)
5. Farooq, U., et al.: Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes. J. Supercomputing 1–34 (2024)
6. Harris, C.J., Krajcik, J.S., Pellegrino, J.W.: Creating and using instructionally supportive assessments in NGSS classrooms. NSTA Press (2024)
7. He, P., Shin, N., Zhai, X., Krajcik, J.: Guiding teacher use of artificial intelligence-based knowledge-in-use assessment to improve instructional decisions: A conceptual framework. In: Zhai, X., Krajcik, J. (eds.) Uses of Artificial Intelligence in STEM Education, pp. xx–xx. Oxford University Press (2024)
8. Hridi, A.P., Sahay, R., Hosseinalipour, S., Akram, B.: Revolutionizing AI-assisted education with federated learning: a pathway to distributed, privacy-preserving, and debiased learning ecosystems. In: Proceedings of the AAAI Symposium Series. vol. 3, pp. 297–303 (2024)
9. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
10. Huang, L.: Ethics of artificial intelligence in education: student privacy and data protection. Sci. Insights Educ. Front. **16**(2), 2577–2587 (2023)
11. Huang, Y., et al.: Personalized cross-silo federated learning on Non-IID data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 7865–7873 (2021)
12. Mistry, D., Mridha, M.F., Safran, M., Alfarhood, S., Saha, A.K., Che, D.: Privacy-preserving on-screen activity tracking and classification in e-learning using federated learning. IEEE Access (2023)
13. PASTA, P.T.: Supporting instructional decision making: Potential of an automatically scored three-dimensional assessment system. https:/ai4stem.org/pasta/ (2023)
14. Porras, J.M., Lara, J.A., Romero, C., Ventura, S.: A case-study comparison of machine learning approaches for predicting student's dropout from multiple online educational entities. Algorithms **16**(12), 554 (2023)
15. Rousi, R., Alanen, H.K., Wilson, A.S.: Data privacy, ethics and education in the era of AI–a university student perspective. In: Proceedings of the Conference on Technology Ethics 2024 (Tethics 2024). RWTH Aachen (2024)
16. Truex, S., et al.: A hybrid approach to privacy-preserving federated learning. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, pp. 1–11 (2019)
17. Xu, B., Yan, S., Li, S., Du, Y.: A federated transfer learning framework based on heterogeneous domain adaptation for students' grades classification. Appl. Sci. **12**(21), 10711 (2022)

18. Zhai, X., He, P., Krajcik, J.: Applying machine learning to automatically assess scientific models. J. Res. Sci. Teach. **59**(10), 1765–1794 (2022)
19. Zhang, P., Zeng, G., Wang, T., Lu, W.: TinyLlama: An open-source small language model. arXiv preprint arXiv:2401.02385 (2024)
20. Zhang, T., et al.: Enhancing dropout prediction in distributed educational data using learning pattern awareness: a federated learning approach. Mathematics **11**(24), 4977 (2023)
21. Zheng, X., Cai, Z.: Privacy-preserved data sharing towards multiple parties in industrial IoTs. IEEE J. Sel. Areas Commun. **38**(5), 968–979 (2020)