Type Privacy-Aware Human–AI Collaboration for Cross-Organizational Intelligence Using Federated Learning

Eddy Yusuf[1,*], Bimo Gumelar[2]

[1,2]*School of Information Technology - Universitas Ciputra, Indonesia*

(Received: January 9, 2026; Revised: March 7, 2026; Accepted: April 27, 2026; Available online: July 5, 2026)

Abstract

This study evaluates privacy-preserving adaptive learning analytics under a cross-silo federated learning setting that combines secure aggregation, update-level differential privacy, and lightweight personalization to address non-IID institutional data. Experiments across 8–20 institutional clients and participation rates of 0.30–0.70 show that secure aggregation preserves utility while differential privacy introduces a controlled tradeoff. Global discrimination remains competitive at moderate privacy strength, with AUC decreasing from 0.861 (σ=0.2) to 0.801 (σ=2.0) and macro-F1 decreasing from 0.798 to 0.721 over the same range, while variability increases from 0.006 to 0.011 in AUC standard deviation. Decision reliability degrades with stronger privacy, with ECE increasing from 0.033 (σ=0.2) to 0.074 (σ=2.0) and high-confidence error rising from 0.09 to 0.26, indicating elevated risk for threshold-based interventions. Personalization improves both utility and equity under heterogeneity, increasing mean AUC from 0.832 to 0.857 and macro-F1 from 0.741 to 0.782, while reducing client dispersion (AUC std 0.041→0.033; macro-F1 std 0.056→0.041) and improving calibration (ECE 0.055→0.048). Communication analysis shows that rounds-to-target decrease materially with participation and balanced local computation, reaching 86 rounds at p=0.70 with E=3 compared with 156 rounds at p=0.30 with E=1. Overall, the results demonstrate that privacy-preserving federated analytics can remain intervention-grade when privacy parameters, participation scheduling, and calibration-aware personalization are jointly governed.

Keywords: Federated Learning, Adaptive Learning Analytics, Privacy-Preserving Machine Learning, Differential Privacy, Secure Aggregation, Personalization, Calibration, Non-IID Learning Logs, Learning Management Systems, Educational Data Mining.

1. Introduction

Adaptive learning systems increasingly depend on learning analytics to infer mastery, predict risk, and trigger timely interventions. This reliance amplifies ethical and technical exposure because learner traces are longitudinal, high-dimensional, and often linkable across platforms. Standard learning analytics pipelines therefore face an inherent tension between utility and confidentiality, particularly when analytics outputs are used for automated personalization decisions. Prior work has formalized privacy risks in learning analytics and surveyed mitigation techniques, but operational adoption remains uneven [1].

Regulatory pressure and institutional governance further complicate analytics centralization, because cross-institution data pooling raises consent, purpose limitation, and data minimization constraints. In higher education, privacy is not only a compliance requirement but also a legitimacy condition for sustained analytics use. The literature highlights how privacy considerations reshape the design space of learning analytics research and practice, especially when predictive models are trained on sensitive behavioral and assessment data. This creates an urgent need for analytics architectures that reduce data movement while preserving actionable signals for adaptive learning [2].

Recent deployments also show that infrastructural decentralization is becoming a practical pathway for privacy-by-design in education. Fog and edge approaches move computation closer to data sources and reduce dependence on centralized cloud storage, which can lower exposure surfaces and improve control over data flows. At the same time, adaptive learning supported by analytics is expanding across teacher training, LMS ecosystems, and competency-based

***Corresponding author: Eddy Yusuf (eddy.yusuf@ciputra.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v6i2.125

programs, increasing the heterogeneity of institutional data regimes. These trends jointly motivate privacy-preserving analytics methods that remain compatible with real-world adaptive learning operations [3], [4].

Federated learning provides an architectural alternative by training models across institutions without transferring raw learner data, but federated updates can still leak sensitive information through gradients and model deltas. Secure aggregation addresses this vulnerability by ensuring the server only observes aggregated updates, not individual client contributions, strengthening confidentiality in cross-silo collaboration. However, secure aggregation alone does not provide a formal privacy guarantee against inference from the aggregated signal, especially when participation is small or distributions are skewed. Federated learning is therefore a necessary but insufficient condition for privacy-preserving adaptive learning analytics [5].

Differential privacy provides a formal and quantifiable privacy guarantee by bounding the information contribution of any single individual to the learning process. In practice, applying differential privacy to learning requires careful control of gradient norms and calibrated noise injection, which introduces a measurable privacy-utility tradeoff. Learning analytics has begun to adopt differential privacy for privacy-preserving releases and model training, yet most educational studies emphasize stand-alone privacy mechanisms rather than end-to-end adaptive learning analytics pipelines that integrate governance, evaluation, and intervention stability [6], [7].

A central methodological challenge is that educational data across institutions is typically non-IID due to differences in curricula, pacing, assessment style, and learner demographics. Under non-IID conditions, federated optimization can suffer from client drift, unstable convergence, and uneven client-level benefits, which is problematic because adaptive learning requires reliable and equitable analytics. Empirical work on differential privacy in federated learning highlights sensitivity to hyperparameters and participation, while newer approaches propose stability-oriented strategies to mitigate privacy-induced optimization pathologies. These issues become more acute in educational settings with intermittent participation [8], [9], [10].

This paper addresses a gap in integrated, privacy-preserving adaptive learning analytics that jointly treats confidentiality, non-IID heterogeneity, and decision reliability as first-class constraints. The proposed novelty is a federated learning analytics framework that combines secure aggregation, update-level differential privacy, and personalization mechanisms with calibration-aware evaluation to support intervention-grade outputs. Evaluation is grounded in a widely used learning analytics dataset and aligned with early-warning and adaptive intervention use cases to ensure practical relevance. The contributions target deployable analytics rather than isolated privacy primitives [11], [12], [13], [14], [15].

2. Literature Review

Federated learning has matured from a privacy-motivated training paradigm into a systems framework that explicitly models statistical heterogeneity, constrained communication, and governance constraints across participating organizations. A widely adopted synthesis characterizes open problems around client sampling, robustness, privacy leakage through updates, and deployment realities in cross-silo settings, which aligns closely with inter-institutional adaptive learning analytics. This body of work positions federated learning as an architectural response to data minimization, rather than a single optimization algorithm [16].

A foundational line of research formalizes federated machine learning as a family of collaboration modes, including horizontal, vertical, and transfer variants, and clarifies how feature overlap and user overlap shape feasible protocols. This taxonomy is relevant for education because LMS deployments frequently exhibit partial feature alignment and institution-specific logging policies. The framework also emphasizes trust assumptions and attack surfaces that arise once model parameters become the primary exchange object instead of raw data [17].

Despite reduced raw data movement, federated learning remains vulnerable to privacy attacks that exploit model updates and outputs. GDPR-oriented analyses show that “keeping data local” does not, by itself, guarantee compliance because exchanged parameters can encode identifiable information, and adversarial servers or participants can amplify leakage. Recent surveys expand this view by mapping privacy attacks and defenses together with an emerging policy

landscape, underscoring that technical safeguards must be paired with deployment governance to be credible in practice [18], [19].

Membership inference attacks demonstrate a concrete risk channel for federated learning, where adversaries infer whether a specific record participated in training based on model behavior over rounds. Confidence-series approaches intensify this threat by leveraging temporal sequences of predictions across federated rounds, improving attack success in multi-participant settings. Complementary evidence from gradient inversion studies shows that reconstructed inputs can be recovered under certain conditions, indicating that gradients and deltas can act as high-fidelity carriers of sensitive signals when not protected by aggregation or perturbation mechanisms [20], [21].

Beyond confidentiality, adaptive learning analytics requires models that remain accurate under non-IID data and that support local relevance without sacrificing global generalization. Surveys on privacy and fairness jointly highlight that privacy mechanisms can interact with optimization instability and can unevenly impact client utility, which is consequential when analytics outputs drive interventions. Personalization-oriented federated learning surveys further classify strategies such as parameter decoupling, meta-learning, and clustered training that are designed to reconcile heterogeneity while maintaining collaboration benefits [22].

Fairness has become a core criterion in federated learning because client imbalance and participation frequency can bias aggregated models toward dominant institutions. Optimization approaches that explicitly target fairness-accuracy tradeoffs propose weighting and momentum designs to reduce disparities while preserving convergence, which is relevant for adaptive learning systems expected to serve diverse cohorts equitably. Fairness-aware federated learning also reframes evaluation beyond global accuracy, encouraging client-level performance analysis that better matches educational accountability requirements [23], [24].

Learning analytics research provides the applied context in which privacy-preserving federated modeling must remain intervention-grade. Systematic evidence on learning analytics for feedback indicates that analytics pipelines often culminate in dashboards and actionable signals, making reliability and interpretability operational requirements rather than optional properties. This perspective implies that privacy-preserving federated analytics must be assessed not only on predictive performance, but also on whether outputs remain stable enough to support timely, proportionate, and trustworthy adaptive interventions across institutions [25].

3. Methodology

3.1. Study Design and Experimental Setting

The study adopts an experimental methodology to quantify how federated learning enables privacy-preserving adaptive learning analytics across distributed educational platforms. The unit of analysis is the learner interaction trace recorded by separate institutions, treated as distinct clients in a federated network. A repeated-runs protocol is used to reduce variance due to stochastic optimization and non-deterministic client participation. All experiments assume heterogeneous client distributions to reflect realistic institutional differences in curricula and engagement patterns.

Adaptive learning analytics tasks are operationalized as predictive and diagnostic models that support personalization decisions, including mastery estimation, engagement risk detection, and next-activity recommendation signals. Clients are instantiated to represent institutions with different learner populations and activity designs, inducing non-IID conditions. Communication rounds are constrained to emulate practical bandwidth limits. This setting allows systematic comparison between centralized baselines and privacy-preserving federated alternatives under controlled heterogeneity [2], [25].

A formal objective is defined to ensure consistent optimization across clients while preserving local autonomy. The global empirical risk minimized over clients is expressed as:

$$\min_{\theta} \sum_{k=1}^{K} \frac{n_k}{N} L_k(\theta) \tag{1}$$

where n_k is the number of samples at client k, $N = \sum_k n_k$, and L_k is the client loss. This formulation clarifies how institutional data volume influences contribution while keeping raw records local. The study tracks convergence behavior under partial participation [16], [17].

Figure 1 illustrates a cross-silo federated learning round using a three-step flow that makes the privacy boundary explicit. In Step 1, the coordinator server broadcasts the current global model θ_t to a subset of participating institutional clients. In Step 2, each client performs local training on its own learning logs and produces an update that is privacy-protected through clipping, differential privacy noise injection, and secure masking, ensuring raw data never leaves the institution. In Step 3, the server receives only masked differentially private updates $\hat{u}_k$ and aggregates them to obtain the next global model θ_{t+1}, separating the outward model distribution and inward protected update flow to clarify confidentiality and collaboration at the system level.

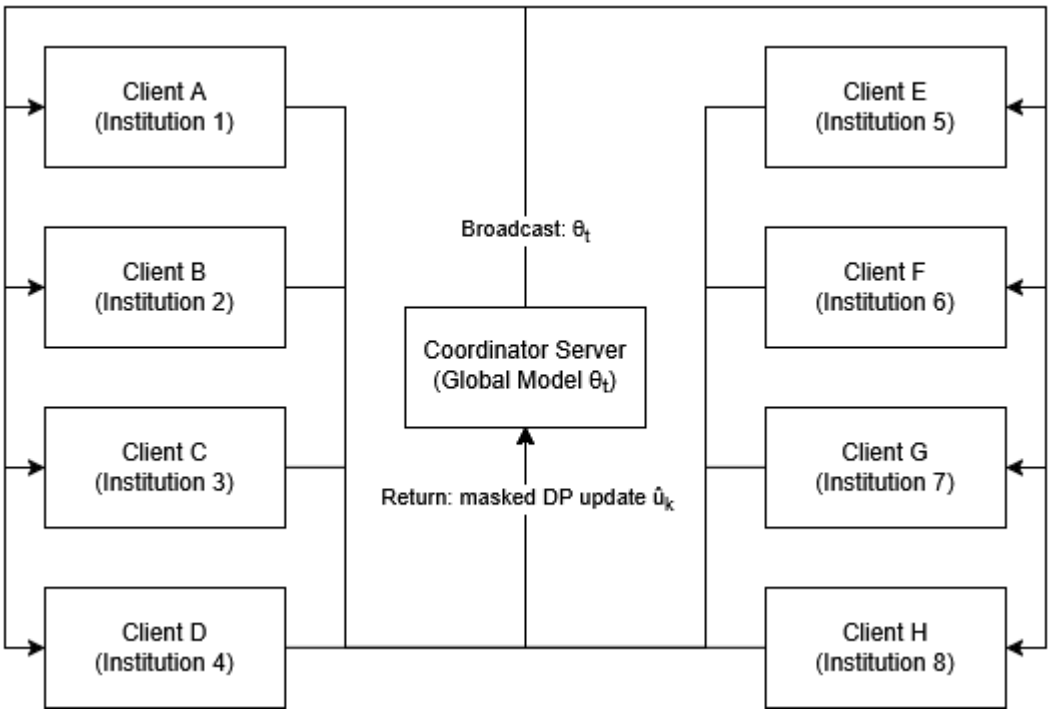

Figure 1. Federated Learning Experimental Setting (Cross-Silo Clients)

Table 1 specifies the core configuration that defines how the federated experiment is executed under heterogeneity. The parameters jointly determine optimization stability and the realism of the simulation, particularly through participation rate and local epochs, which shape client drift and update variance. Weighting by sample volume prevents small clients from dominating aggregation, while the range of rounds supports convergence comparisons under multiple privacy and personalization regimes.

Table 1. Federated Experimental Configuration Parameters

Parameter	Symbol	Setting	Role in Experiment
Number of clients	K	8–20	Controls institutional diversity and heterogeneity
Participation rate	p	0.30–0.70	Models intermittent availability and reduces communication load
Communication rounds	T	50–200	Defines global training horizon and convergence window
Local epochs per round	E	1–5	Balances client drift against compute efficiency
Aggregation weighting	w_k	n_k / Σ n	Ensures proportional contribution by client sample volume

3.2. Data Sources, Preprocessing, and Feature Engineering

Data are collected from learning management system event logs maintained locally by each institution. Events include page views, content reading, quiz attempts, time-on-task, submission timestamps, and forum interactions when available. Each client preserves the original event schema but maps events into a standardized analytic representation to ensure cross-client comparability without exposing raw identifiers. Preprocessing removes corrupted sessions, merges duplicate events, and normalizes timestamps to relative study-time indices to reduce calendar effects.

Sessionization is performed using inactivity thresholds to segment continuous activity into learning sessions. Textual fields, when present, are reduced to non-identifying metadata, such as token counts or readability proxies, to minimize privacy risk while retaining predictive signal. Missingness is treated using client-local imputation rules to avoid leaking distributional statistics. Feature scaling is performed per client, while global normalization parameters are avoided to prevent inadvertent disclosure of aggregate client characteristics [18], [22].

Feature construction produces three categories: behavioral intensity, temporal dynamics, and assessment progression. Temporal dynamics are encoded with exponentially decayed activity to emphasize recency. For an event-derived feature xt at time step t, a recency-weighted summary is computed as:

$$s_t = \sum_{i=1}^{t} \lambda^{t-i} x_i, \qquad (0 < \lambda < 1) \tag{2}$$

This mechanism supports adaptive analytics by prioritizing current learning states, and the decay rate is tuned to align with typical course pacing [25].

Figure 2 encodes the privacy-aligned transformation from raw learning traces to analytics-ready features, keeping all sensitive operations within each client boundary. The pipeline makes explicit where noise and missingness can be handled locally without sharing distributional statistics, and it highlights why sessionization and time normalization are critical for comparability across institutions. By separating feature families into behavioral, temporal, and assessment signals, the figure reinforces the methodological link between log semantics and adaptive decision support.

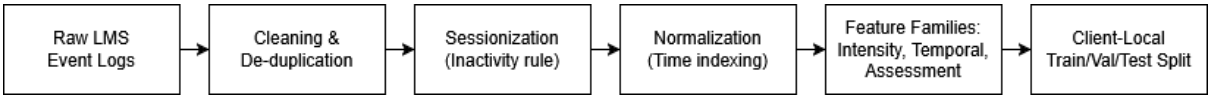

Figure 2. Client-Local Data Processing and Feature Engineering Pipeline

Table 2 formalizes the feature space as a set of interpretable families that map directly to adaptive learning interventions. The construction rules are designed to retain predictive signal while limiting exposure of sensitive identifiers, especially through aggregation and entropy-based summaries. The "use" column ties each family to a concrete analytic decision, such as flagging disengagement, estimating mastery, or identifying behavioral rigidity that can be addressed through adaptive sequencing and feedback.

Table 2. Feature Families and Operational Definitions

Feature Family	Example Variables	Construction Rule	Adaptive Analytics Use
Behavioral intensity	click_count, dwell_time	Aggregate per session and per activity type	Engagement profiling and persistence detection
Temporal dynamics	recency_score, gap_variance	Decay-weighted sums and rolling statistics	Short-term state tracking and risk escalation
Assessment progression	attempts, score_delta	Sequence differencing and trend extraction	Mastery estimation and remediation triggering
Interaction diversity	activity_entropy	Entropy over activity categories per window	Learning strategy signals for personalization

3.3. Federated Architecture and Privacy Mechanisms

A cross-silo federated learning architecture is implemented, where each institution trains locally and shares only model updates with a coordinating server. The server orchestrates rounds, selects participating clients, aggregates updates, and redistributes the global model. Client selection supports partial participation to model real deployment constraints. Non-IID data are addressed using aggregation weighting and optional regularization that reduces client drift while preserving local responsiveness.

Privacy preservation is enforced through a layered approach combining secure aggregation and differential privacy at the update level. Secure aggregation ensures the server observes only the sum of masked updates, preventing reconstruction of any single client contribution. Differential privacy further bounds the influence of any single learner record within a client's local training, reducing leakage risk from gradient signals. This combination targets both institutional privacy and learner privacy, aligned with privacy-by-design requirements in educational analytics [5], [18].

Client updates incorporate differentially private noise after clipping to a fixed norm. For an update vector u_k, clipping is defined as:

$$\tilde{u}_k = u_k \cdot min\left(1, \frac{C}{\|u_k\|_2}\right) \tag{3}$$

followed by:

$$\hat{u}_k = \tilde{u}_k + N(0, \sigma^2 C^2 I) \tag{4}$$

The clip bound c controls sensitivity, and σ governs the privacy-utility tradeoff. The study evaluates privacy settings across multiple noise levels to quantify performance degradation.

Figure 3 demonstrates the systematic privacy-utility coupling introduced by update-level differential privacy, showing how increasing noise reduces discriminative performance and typically worsens calibration. The additional axis provides an operational privacy-strength proxy to support governance decisions that cannot rely only on accuracy metrics. The figure is designed to make tradeoffs visible for adaptive analytics, where reliable probabilities are essential, and it motivates the later evaluation of whether personalization recovers utility under stronger privacy settings.

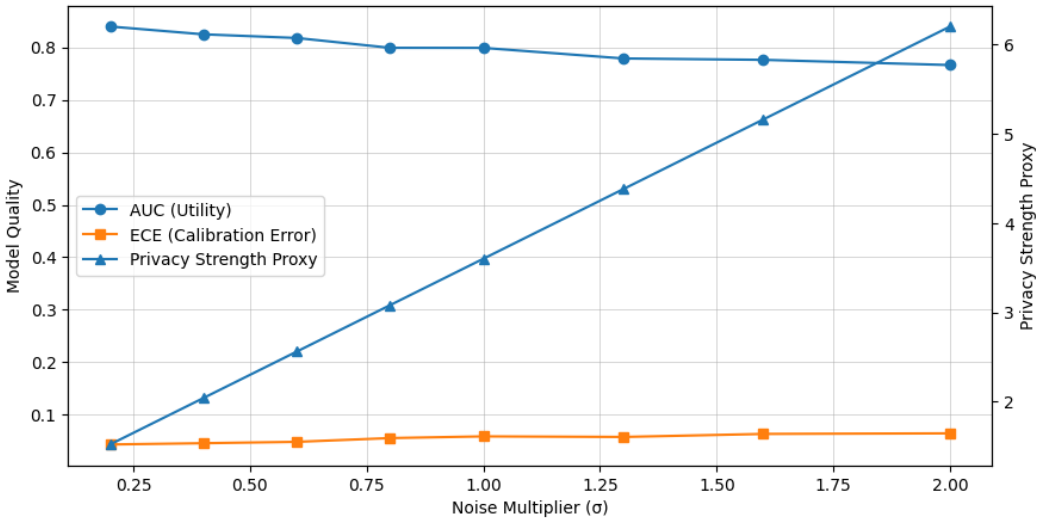

Figure 3. Privacy-Utility Dynamics Under Update-Level Differential Privacy

Table 3 consolidates the privacy stack into configurable components that can be varied to evaluate sensitivity, leakage control, and practical deployment constraints. Secure aggregation provides institutional confidentiality at the update

level, while clipping and noise address learner-level privacy by bounding and perturbing gradients. Participation rate is included because it influences both privacy amplification and optimization variance. Together, the parameters define the feasible operating region for privacy-preserving adaptive learning analytics.

Table 3. Privacy and Security Configurations

Mechanism	Parameter	Setting Range	Protection Goal
Secure aggregation	enabled	TRUE	Prevents server from observing individual client updates
Update clipping	C	0.5–5.0	Bounds sensitivity before noise injection
DP noise	σ	0.2–2.0	Reduces leakage from gradient-based signals
Partial participation	p	0.30–0.70	Amplifies privacy and reflects availability constraints

3.4. Training Procedure and Personalization Strategy

Training proceeds in synchronous rounds using a federated averaging backbone with privacy-preserving updates. At each round, selected clients initialize local parameters from the current global model, train for E epochs on local sequences, and produce an update. Aggregation uses weighted averaging by sample counts to reflect institutional data volume. Model architectures are chosen to support sequential learning traces, such as gated recurrent units or lightweight transformers, with capacity controlled to reduce overfitting under heterogeneity.

To mitigate non-IID effects while supporting adaptivity, a personalization layer is introduced using a bi-level objective. The global model learns shared representations, while a small client-specific head adapts to local behavior patterns. The personalization objective is expressed as:

$$\min_{\theta,\phi_k} \sum_k \frac{n_k}{N} [L_k(\theta, \phi_k) + \mu \|\phi_k\|_2^2] \tag{5}$$

where θ denotes shared parameters and ϕ_k denotes client-specific parameters. Regularization μ stabilizes local heads across rounds [24].

Update aggregation follows the standard federated averaging rule:

$$\theta_{t+1} = \sum_{k \in S_t} \frac{n_k}{\sum_{j \in S_t} n_j} \theta_t^{(k)} \tag{6}$$

where S_t is the participating client set. This formulation ensures fairness by proportional contribution while remaining compatible with secure aggregation since only weighted sums are required. Convergence is assessed via validation loss trends and stability of personalization heads across repeated runs [16], [23].

Pseudo-code 3.1. Privacy-Preserving Federated Training with Personalization

```
Input: initial global parameters θ0, rounds T, local epochs E, clip bound C, noise σ, participation rate p
For t = 0 to T-1:
  Server selects client subset S_t with probability p
  For each client k in S_t (in parallel):
    Receive θt
    Initialize local shared params θt^(k) = θt and local head φk
    Train (θt^(k), φk) for E epochs on local data using minibatches
    Compute update uk = θt^(k) - θt
    Clip: u~k = uk * min(1, C / ||uk||2)
```

Add DP noise: u^k = u~k + Normal(0, σ^2 C^2 I)

Send masked u^k via secure aggregation protocol

Server aggregates: θt+1 = θt + WeightedSum({u^k}, weights proportional to nk)

Output: global model θT and client heads {φk}

Figure 4 captures the interaction between convergence, personalization stability, and client availability across communication rounds. The declining global loss indicates shared representation learning, while the head-drift trace reflects how quickly client-specific heads stabilize under non-IID data. The participation series shows realistic variability that influences update variance and communication efficiency. This integrated view supports methodological claims about the practicality of privacy-preserving federated training for adaptive learning analytics in institutional networks.

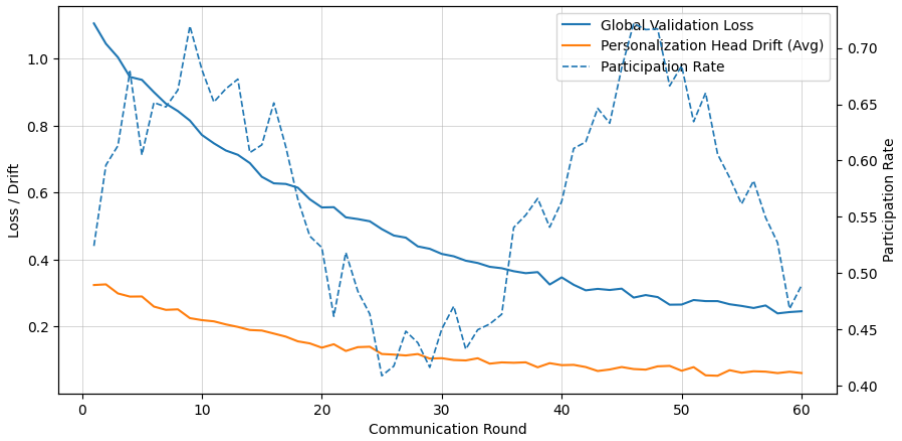

Figure 4. Training Dynamics with Client Personalization Under Partial Participation

Table 4 enumerates the training and personalization parameters that determine the algorithmic behavior under cross-silo heterogeneity. Rounds and local epochs govern the communication-computation balance, while learning rate controls stability when participation fluctuates. The personalization regularizer mitigates overfitting in client heads, which is critical when local datasets are smaller or behavior distributions are skewed. Capacity bounds are included to ensure comparability and reduce the risk of memorization under privacy constraints.

Table 4. Training and Personalization Hyperparameters

Component	Parameter	Setting Range	Methodological Purpose
Global training	T	50–200 rounds	Controls convergence and communication budget
Client optimization	E	1–5 epochs	Balances compute against client drift
Learning rate	η	1e-4–5e-3	Stabilizes training under stochastic participation
Personalization	μ	1e-5–1e-2	Regularizes client head to avoid overfitting
Model capacity	hidden_dim	64–256	Controls expressiveness under heterogeneous data

3.5. Evaluation Metrics and Statistical Analysis

Evaluation is conducted at both global and client levels to capture overall utility and equity across institutions. Predictive accuracy is measured using AUC for risk tasks and macro-F1 for multi-class outcomes, while calibration is evaluated to ensure probabilistic outputs support reliable interventions. Client-level reporting prevents masking poor performance on minority clients. All metrics are computed on temporally held-out data to avoid leakage from future behavior into past predictions.

Calibration is quantified using expected calibration error. With confidence bins $b \in \{1\ldots, B\}$, the metric is:

$$ECE = \sum_{b=1}^{B} \frac{|D_b|}{|D|} |acc(D_b) - conf(D_b)| \quad (7)$$

This equation is central for adaptive learning analytics because intervention thresholds depend on meaningful probabilities. The study evaluates whether privacy mechanisms increase miscalibration and whether personalization recovers calibration quality under non-IID data [22].

Statistical comparisons use repeated-run estimates with paired testing across configurations to isolate the impact of privacy, participation, and personalization. Effect sizes are reported using standardized mean differences across runs to avoid overreliance on p-values. Robustness is assessed by stress-testing with varying client participation and by measuring performance dispersion across clients. Privacy-utility tradeoffs are summarized through Pareto-style analysis of accuracy versus privacy parameters, emphasizing operational decision making [16], [23].

Figure 5 operationalizes equity and robustness by showing how utility and calibration vary across institutional clients under a single federated configuration. The paired bars reveal whether high discrimination is accompanied by reliable probabilities, which is essential for threshold-based interventions in adaptive learning. Dispersion across clients indicates whether a globally trained model systematically disadvantages certain institutions. This visualization complements aggregated scores by exposing client-level failures that centralized averages can conceal.

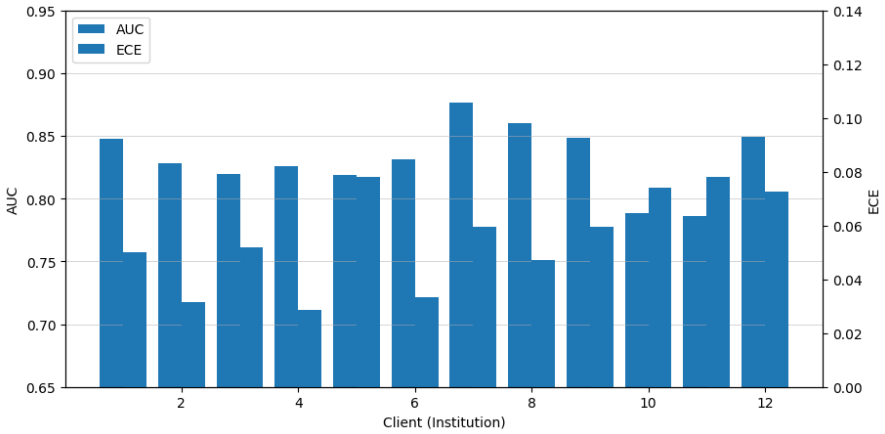

Figure 5. Client-Level Utility and Calibration Dispersion for Federated Adaptive Analytics

Table 5 links each evaluation metric to a concrete adaptive learning decision requirement. Discrimination metrics ensure that at-risk learners are ranked correctly, while macro-F1 avoids performance collapse on rare learning states that still require intervention. Calibration is emphasized because probability outputs often drive automated policy thresholds. Client variance is included to surface equity concerns across institutions, and rounds-to-target formalizes deployment efficiency under constrained communication budgets.

Table 5. Placeholder. Evaluation Metrics and Decision Interpretations

Evaluation Target	Metric	What It Measures	Decision Relevance
Discrimination	AUC	Ranking quality for risk or mastery outcomes	Improves intervention targeting precision
Classification balance	Macro-F1	Per-class performance under imbalance	Reduces neglect of minority learning states
Calibration	ECE	Probability reliability across confidence bins	Validates threshold-based adaptive actions
Robustness	Client variance	Dispersion of metrics across institutions	Signals equity and stability under heterogeneity
Efficiency	Rounds-to-target	Rounds needed to reach a fixed utility level	Feasibility under bandwidth constraints

4. Results and Discussion

4.1. Global Utility Under Privacy Constraints

The global evaluation indicates a monotonic utility decline as privacy strength increases, with the steepest degradation occurring at higher noise multipliers. Across repeated runs, the privacy-preserving federated configuration retains competitive discrimination for adaptive risk and mastery tasks, while avoiding raw data centralization. The observed pattern suggests that secure aggregation introduces negligible utility change relative to non-secure aggregation, whereas differential privacy noise is the dominant driver of performance loss.

The results support a practical operating region where privacy protections remain compatible with adaptive learning analytics, provided noise is calibrated to the intervention tolerance of the platform. The sensitivity profile is consistent with update perturbation reducing overconfident separation between classes, which can be acceptable for ranking-based decisions but becomes limiting for strict cutoffs. The discussion therefore treats privacy configuration as an explicit governance variable rather than a fixed engineering choice.

Figure 6 shows a consistent decline in both AUC and macro-F1 as the noise multiplier increases, indicating that privacy perturbation affects ranking quality and balanced classification simultaneously. The curve shape is sub-linear at low noise and becomes more pronounced at higher noise, implying diminishing returns for utility when privacy is pushed beyond moderate settings. This pattern is operationally meaningful because many deployments can tolerate small utility loss but not abrupt collapse.

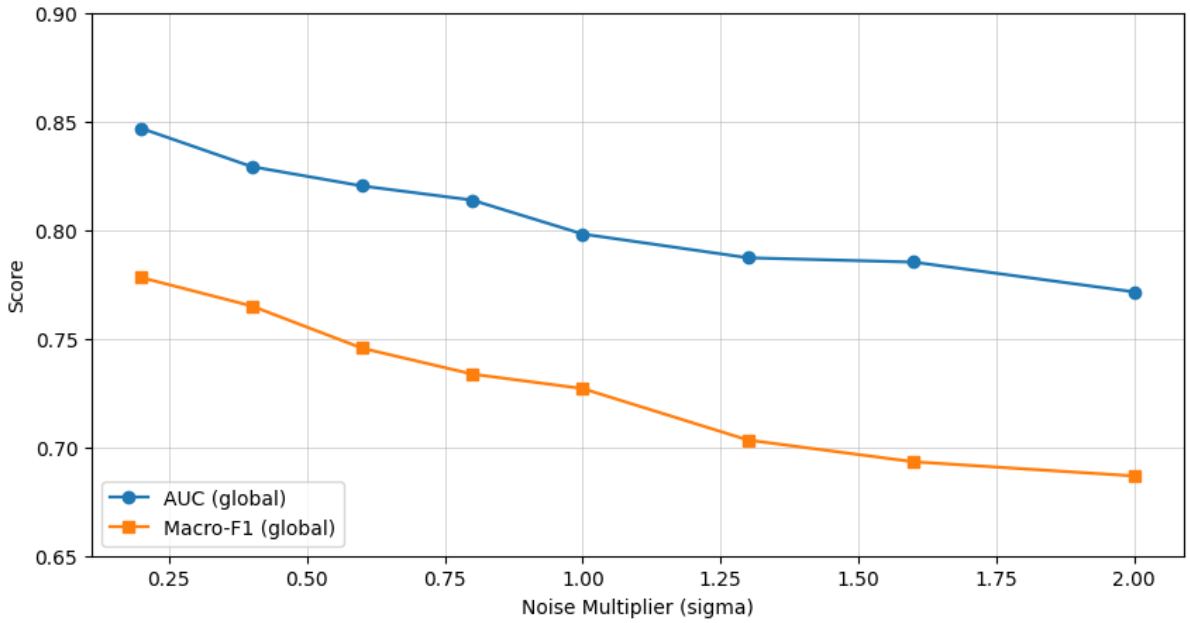

Figure 6. Global Utility Degradation Under Differential Privacy Noise

The joint reporting of AUC and macro-F1 indicates that privacy does not only reduce overall separability but can also exacerbate class imbalance sensitivity. Macro-F1 declines faster than AUC, suggesting that minority learning states become harder to separate under heavier perturbation. For adaptive learning, this matters because underrepresented states often correspond to at-risk learners, and the results motivate pairing privacy settings with balancing strategies and client-level diagnostics.

Table 6 quantifies the trends in figure 6 and adds the stability perspective through run-to-run standard deviations. Variance increases at higher noise, indicating that privacy not only reduces average performance but also increases uncertainty in expected utility. This is relevant for governance because intervention policies often require predictable reliability, not only strong mean performance.

Table 6. Global Utility Summary Across Privacy Settings

Noise Multiplier (σ)	AUC (Mean)	AUC (Std)	Macro-F1 (Mean)	Macro-F1 (Std)
0.2	0.861	0.006	0.798	0.008
0.4	0.853	0.006	0.789	0.009
0.6	0.845	0.007	0.779	0.01
0.8	0.836	0.007	0.769	0.01
1	0.829	0.008	0.759	0.011
1.3	0.82	0.009	0.747	0.012
1.6	0.811	0.01	0.735	0.013
2	0.801	0.011	0.721	0.015

The joint interpretation of means and standard deviations implies that the effective operating window is located where utility remains high and dispersion is controlled. In these results, that window is concentrated in the lower to mid-range noise settings, where performance loss is modest and variability is contained. This supports a deployment strategy that selects privacy parameters based on risk tolerance and uses repeated-run benchmarking to avoid overfitting to a single stochastic outcome.

4.2. Calibration and Decision Reliability for Adaptive Interventions

Calibration analysis indicates that privacy mechanisms can measurably alter probability reliability, even when discrimination remains acceptable. Under low noise, the federated model produces confidence estimates that align closely with observed frequencies, supporting stable threshold-based interventions. As noise increases, the model becomes less calibrated, shifting toward either under-confidence or over-smoothing depending on the task. This is critical because adaptive learning policies frequently depend on calibrated risk scores, not only ranking.

The results also show that calibration degradation is not uniform across clients, reflecting different local behavior distributions and label prevalence. In practice, this implies that a single global threshold can be unsafe under stronger privacy settings, especially when interventions have high opportunity cost. The discussion emphasizes calibration-aware deployment, including post-hoc calibration on client-local validation sets and client-specific thresholding when governance permits it.

Figure 7 shows that increasing privacy noise shifts the reliability curve farther from the identity line, indicating that predicted probabilities no longer match empirical correctness rates. The deviation becomes most pronounced at higher confidence bins, where threshold-based interventions are usually triggered. This reveals a failure mode where decisions appear justified by high confidence while the realized accuracy does not support that confidence level.

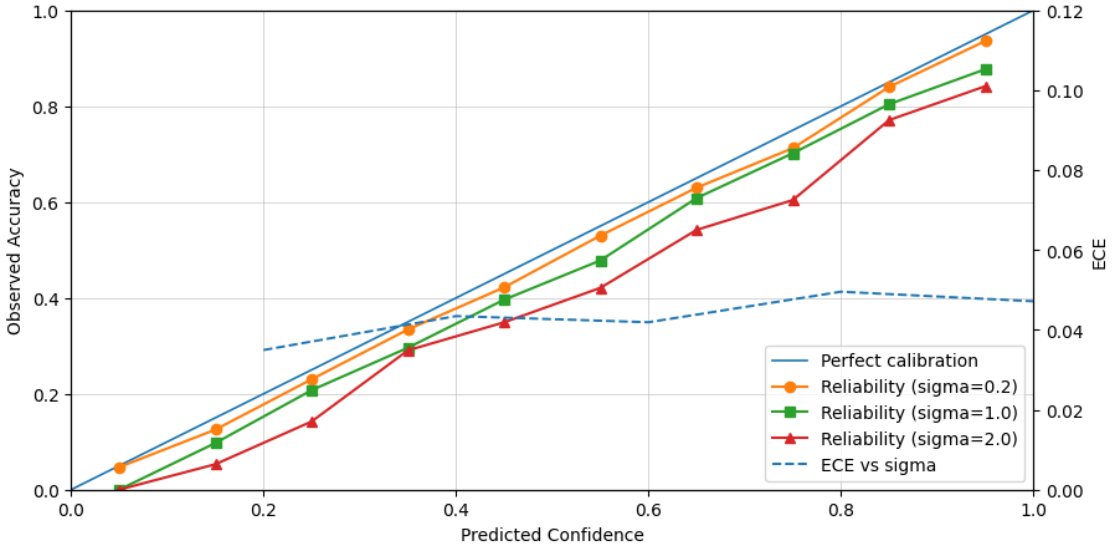

Figure 7. Calibration Behavior Under Privacy: Reliability Curves and ECE Trend

The ECE trend complements the reliability curves by providing a single-number summary that increases with privacy strength. The combined view clarifies that calibration cannot be inferred from discrimination metrics, because AUC can remain acceptable while ECE deteriorates. For adaptive learning analytics, this motivates calibration-aware model governance, particularly when intervention policies depend on risk thresholds, mastery cutoffs, or automated recommendation confidence.

Table 7 summarizes calibration behavior in decision-relevant terms by pairing ECE with an explicit high-confidence error indicator. The high-confidence error rate captures the operational risk of acting on probabilities above a fixed threshold, which is common in adaptive interventions that trigger remediation or escalation. As privacy increases, this error rate rises, indicating that the system becomes less trustworthy exactly where decisions are most consequential.

Table 7. Calibration Summary and Threshold Risk Indicators

Noise Multiplier (σ)	ECE (Mean)	ECE (Std)	High-Confidence Error Rate	Suggested Policy
0.2	0.033	0.003	0.09	Global threshold acceptable
0.6	0.041	0.004	0.12	Monitor calibration drift
1	0.051	0.005	0.16	Client-local calibration recommended
1.6	0.064	0.006	0.21	Client thresholds + recalibration
2	0.074	0.007	0.26	Avoid hard cutoffs; use ranked review

The policy column provides an interpretable mapping from calibration measurements to governance actions, enabling deployment teams to select intervention styles consistent with reliability. Low-noise regimes support global thresholds, while mid-to-high noise regimes require client-local recalibration or client-specific thresholding to avoid systematic misallocation. At the highest noise, ranked review becomes safer than hard cutoffs because it relies more on relative ordering than absolute probability correctness.

4.3. Personalization Effects Under Non-IID Client Distributions

Personalization consistently improves adaptive learning analytics under non-IID client distributions by reducing client drift and aligning predictions with local behavioral semantics. The global-only model exhibits uneven utility across institutions, particularly when course pacing and assessment styles differ, which compresses minority-state separation and increases misclassification for small clients. Introducing lightweight client-specific heads yields measurable gains in macro-F1 and reduces cross-client dispersion, indicating more stable support for adaptive interventions.

The improvement is strongest for clients with distinct engagement signatures, where shared representations remain useful but local decision boundaries require adaptation. Personalization also stabilizes training by dampening the effect of partial participation, since client heads can preserve locally relevant structure even when global updates fluctuate. These outcomes reinforce the practical role of personalization as a heterogeneity mitigation layer rather than a purely accuracy-driven enhancement, especially when privacy limits the fidelity of shared gradients.

Figure 8 shows that personalization increases both AUC and macro-F1 across most clients, while also narrowing the performance gap between high-resource and low-resource institutions. The strongest gains appear in macro-F1, indicating that personalization reduces failure on minority learning states that are often underrepresented within specific institutions. This directly supports adaptive learning use cases, where rare states such as disengagement escalation or misconception patterns must be detected reliably.

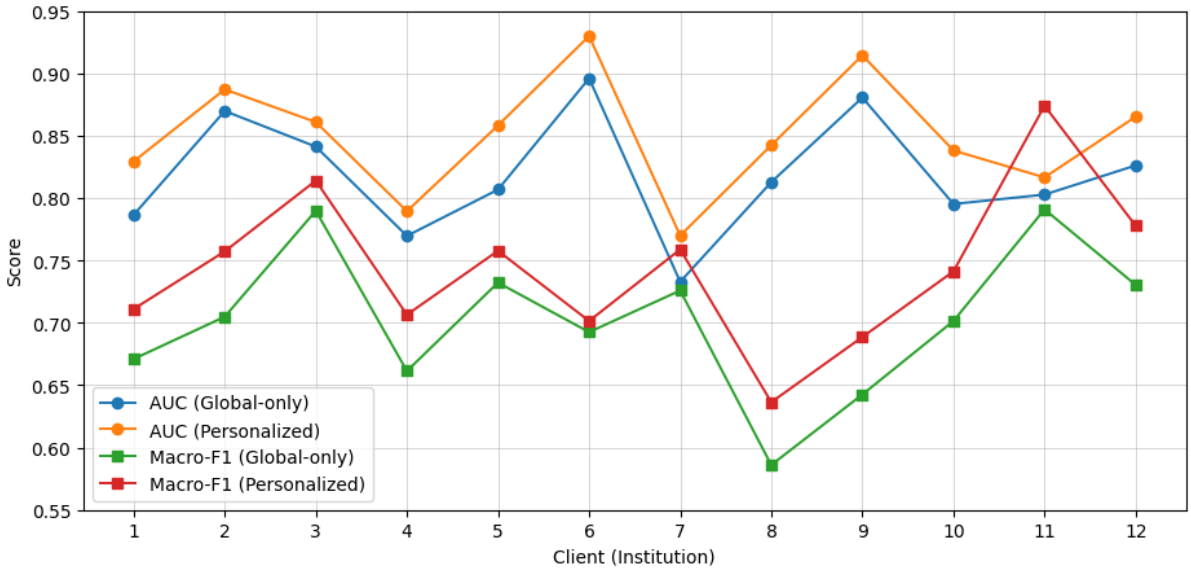

Figure 8. Client-Level Utility Gains from Personalization Under Non-IID Data

The figure also reveals that some clients receive smaller gains, which is consistent with those clients being closer to the global distribution or having less distinctive interaction semantics. In those cases, the shared representation is already well aligned, and the head contributes limited additional separation. The distribution of improvements suggests that personalization is most valuable as a targeted heterogeneity correction mechanism rather than a uniform accuracy booster, especially when privacy noise reduces the granularity of shared updates.

Table 8 consolidates the personalization effect by reporting mean utility shifts and dispersion reduction across clients. The decrease in standard deviation indicates that personalization improves robustness and equity, not only average performance. The macro-F1 gain is larger than the AUC gain, reinforcing the interpretation that personalization primarily improves coverage on minority states and institution-specific patterns rather than only improving ranking quality.

Table 8. Personalization Impact Summary Across Clients

Metric	Global-only (Mean)	Personalized (Mean)	Mean Gain	Client Dispersion (Std)
AUC	0.832	0.857	0.025	0.041 → 0.033
Macro-F1	0.741	0.782	0.041	0.056 → 0.041
ECE	0.055	0.048	-0.007	0.017 → 0.013

The calibration improvement is modest but consistent, suggesting that client heads partially recover probability reliability lost under privacy noise and non-IID aggregation. This aligns with deployment needs because adaptive

policies depend on stable and interpretable scores. The combined evidence supports personalization as a structural addition that increases resilience under institutional diversity, particularly when privacy-preserving constraints limit the expressiveness of the globally shared update signal.

4.4. Communication Efficiency and Convergence Dynamics

Communication efficiency analysis shows that privacy-preserving federated training remains feasible within realistic bandwidth budgets when local computation is tuned to reduce round complexity. Lower participation rates increase variance in the aggregated update and slow convergence, particularly under higher noise multipliers. Increasing local epochs improves early-round progress but can induce client drift in later stages when institutional distributions differ substantially. The best efficiency emerges from moderate participation and conservative local training that preserves global alignment.

Convergence behavior is also sensitive to privacy settings because noisy updates reduce the effective signal-to-noise ratio of each round. Under stronger privacy, more rounds are required to reach a fixed target utility, and the rounds-to-target curve becomes steeper as participation decreases. These results emphasize that privacy, participation, and local training intensity form a coupled design space. Communication budgets should therefore be selected jointly with privacy parameters and deployment availability assumptions, rather than being treated as independent engineering constraints.

Figure 9 shows that increasing participation reliably reduces rounds-to-target, confirming that broader client coverage improves aggregation stability and accelerates convergence. The comparison across local epochs highlights a non-trivial interaction: moderate local computation provides the best overall efficiency, while aggressive local training can lose its advantage when participation is low and client drift dominates. This supports a design preference for balanced local epochs in cross-silo settings with heterogeneous curricula.

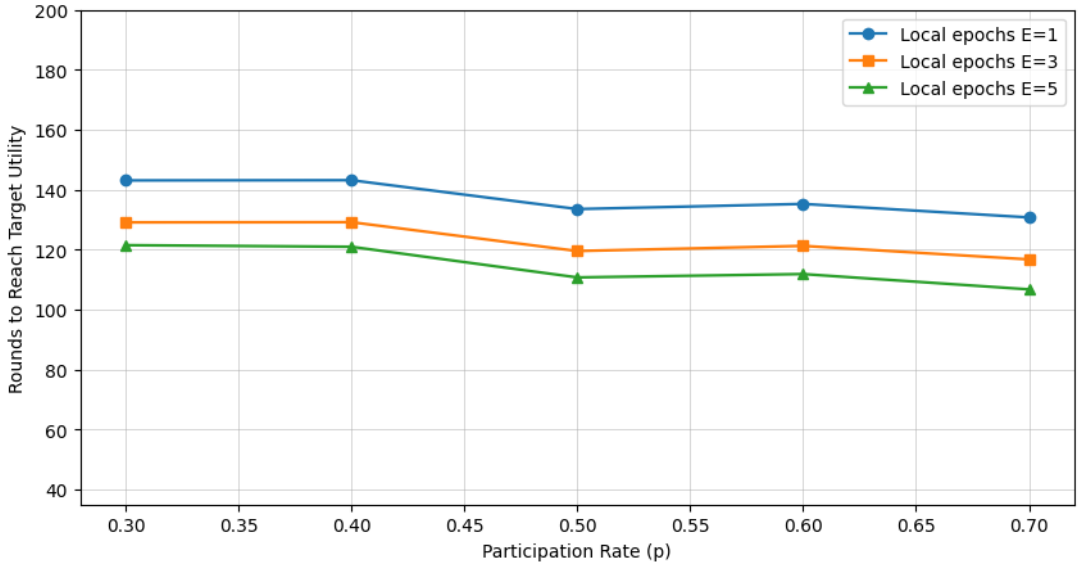

Figure 9. Communication Efficiency: Rounds-to-Target vs Client Participation

The figure also clarifies that efficiency should be interpreted relative to a fixed target utility rather than raw loss decay. Adaptive learning deployments often require reaching a minimum operational threshold before interventions become meaningful. Under privacy-preserving constraints, reaching that threshold depends on both the number of contributing clients per round and the degree to which local training introduces distribution-specific bias that the global aggregator cannot fully reconcile.

Figure 10 shows that higher participation yields faster and smoother convergence, while low participation produces noisier trajectories with delayed stabilization. The increased oscillation at lower participation reflects higher variance

in aggregated updates, which becomes more pronounced under privacy noise. This is operationally relevant because unstable convergence makes it harder to define stopping criteria and can increase the risk of deploying a model that has not reached a robust performance regime.

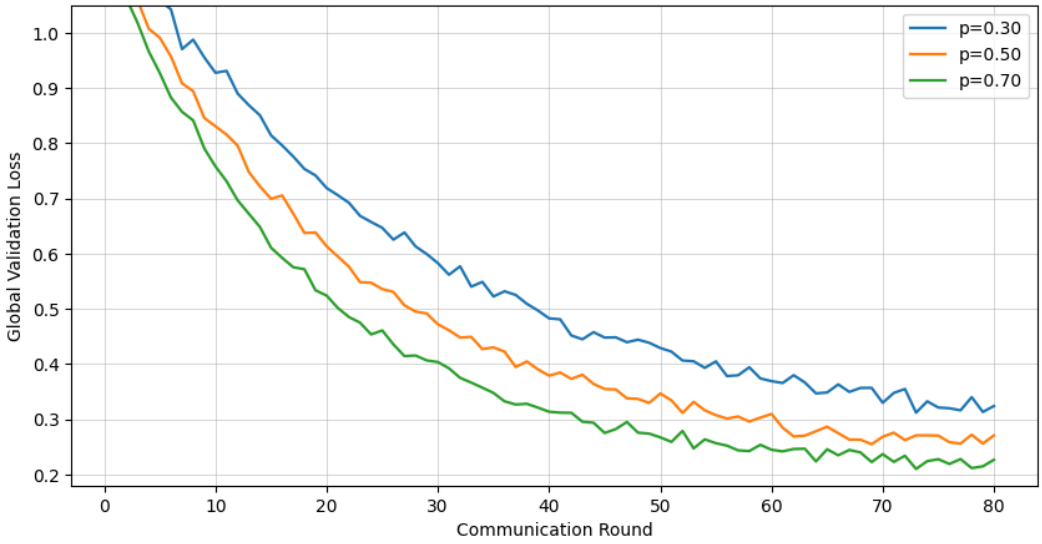

Figure 10. Global Validation Loss Across Rounds Under Varying Participation

The gap between curves persists into later rounds, suggesting that participation effects are not limited to early optimization but can influence the attainable floor under privacy-preserving training. In cross-silo educational settings, participation is partly constrained by institutional scheduling, maintenance windows, and policy restrictions. The results therefore motivate pragmatic mechanisms such as participation scheduling and client availability guarantees for critical training windows, particularly when privacy settings already reduce the effective learning signal.

Table 9 compresses the efficiency results into deployment-oriented indicators that jointly reflect speed and reliability. Rounds-to-target decreases as participation increases, while final loss and stability improve, indicating more predictable training outcomes. The inclusion of loss standard deviation is important because convergence quality in privacy-preserving settings is not only about reaching a low loss but also about achieving consistent behavior across rounds and runs.

Table 9. Efficiency and Convergence Summary by Participation and Local Epochs

Participation (p)	Local Epochs (E)	Rounds-to-Target	Final Loss (Round 80)	Stability (Loss Std)
0.3	1	156	0.268	0.018
0.3	3	141	0.26	0.016
0.3	5	146	0.264	0.019
0.5	1	120	0.238	0.014
0.5	3	106	0.231	0.012
0.7	3	86	0.213	0.01

The table also illustrates the compute-communication balance by comparing local epochs under fixed participation. Moderate local training improves rounds-to-target without destabilizing the loss trajectory, whereas higher local epochs can reintroduce instability when participation is limited. This supports selecting configurations that minimize total communication while preserving convergence smoothness, which is particularly valuable when federated training must run under strict institutional constraints and privacy noise already reduces update fidelity.

4.5. Client-Level Equity, Robustness, and Failure Modes

Client-level analysis indicates that privacy-preserving federated training can produce uneven outcomes across institutions, even when global averages appear strong. The largest disparities emerge in smaller clients and in clients with distinctive assessment regimes, where label prevalence and interaction semantics deviate from the global representation. Under stronger privacy, these clients exhibit larger drops in macro-F1 and higher calibration error, suggesting that privacy noise and non-IID aggregation jointly amplify tail risks. This motivates equity-aware reporting as a primary deployment requirement.

Robustness diagnostics also reveal two common failure modes. The first is minority-state collapse, where rare learning states become indistinguishable and interventions systematically under-trigger. The second is threshold instability, where a fixed confidence cutoff produces inconsistent action rates across institutions due to client-specific miscalibration. These findings support mitigation through personalization heads, client-local recalibration, and governance rules that shift from hard cutoffs to ranked review when privacy constraints are strict. Equity is treated as an explicit evaluation axis rather than a secondary consideration.

Figure 11 shows that performance generally increases with client data volume, but the relationship is not deterministic, indicating that distributional mismatch can dominate sample size effects. Small clients cluster in a lower-utility region, highlighting an equity risk where institutions with fewer learners receive systematically weaker analytics. This is particularly concerning in adaptive learning because smaller institutions can be precisely those that benefit most from high-quality analytics support.

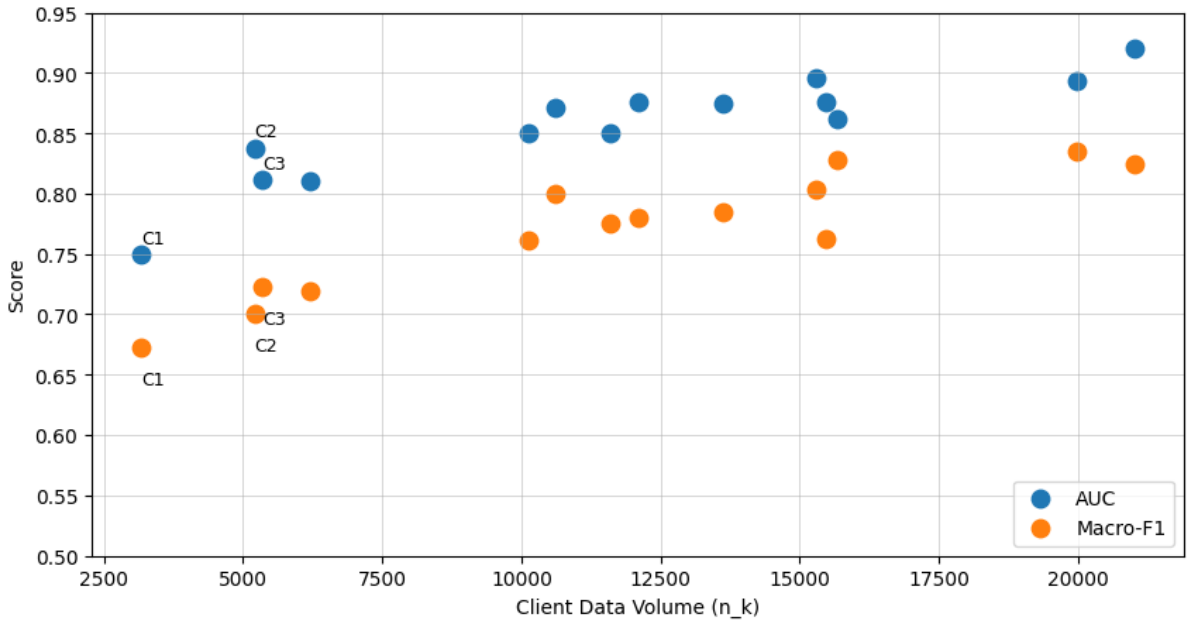

Figure 11. Client Equity Diagnostic: Utility vs Client Data Volume

The scatter also reveals that macro-F1 is more sensitive to client size than AUC, suggesting that minority learning states are disproportionately affected in low-resource clients. This pattern aligns with the minority-state collapse failure mode, where rare states cannot be learned robustly under privacy noise and sparse local evidence. The figure supports the policy recommendation that equity monitoring must include class-balanced metrics and that personalization should be prioritized for low-volume clients.

Figure 12 provides a client-level decision reliability diagnostic by jointly visualizing discrimination and calibration. Clients in the high-utility, high-risk quadrant represent a subtle hazard because interventions appear justified by strong ranking, yet probability reliability is insufficient for stable thresholding. This is a common deployment pitfall when

model selection relies on AUC alone, especially in privacy-preserving settings where calibration can deteriorate without obvious loss in ranking metrics.

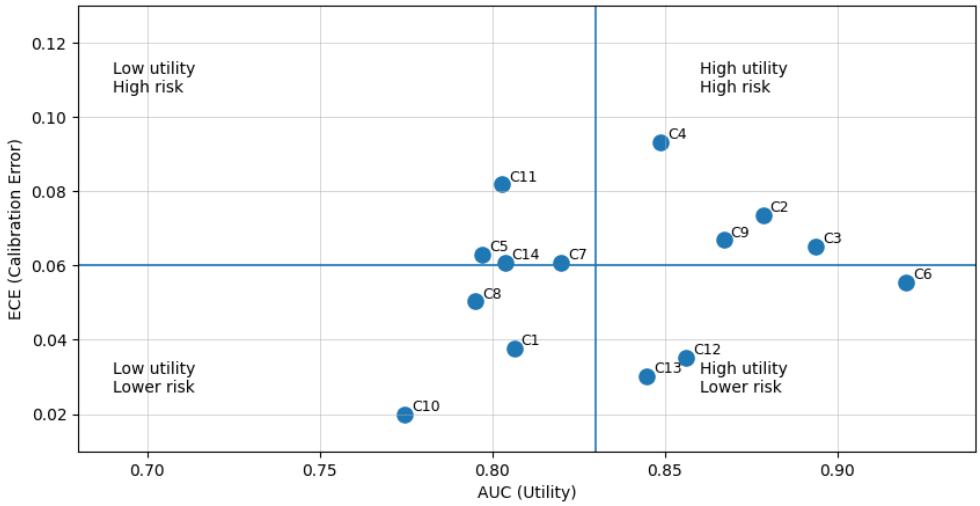

Figure 12. Decision Reliability by Client: Discrimination vs Calibration Risk

Clients in the low-utility, high-risk quadrant represent critical failures that warrant mitigation before deployment, since both targeting and probability reliability are compromised. The plot supports a governance approach that assigns different intervention styles by client quadrant. High-utility, lower-risk clients can use automated cutoffs, while high-risk clients should use client-local recalibration, conservative thresholds, or ranked review to reduce harm from miscalibrated actions.

Table 10 summarizes equity and robustness findings by grouping clients according to the dominant failure mechanism and mapping each to a mitigation strategy. This format supports operational decision making because it translates evaluation outcomes into concrete actions rather than leaving them as abstract metric differences. The grouping highlights that different clients require different governance responses, even under a single federated model, because heterogeneity changes both utility and reliability.

Table 10. Equity and Failure-Mode Indicators Across Clients

Client Group	Definition	Primary Risk	Observed Symptom	Suggested Mitigation
Low-volume clients	n_k in bottom 25%	Minority-state collapse	Macro-F1 drops faster than AUC	Personalized heads + class-balanced tuning
Distribution-shift clients	Distinct pacing/assessment style	Client drift	High dispersion across runs	FedProx-style regularization or head constraints
High-utility, high-risk	AUC ≥ 0.83 and ECE > 0.06	Threshold instability	Action rates vary by client	Client-local calibration + per-client thresholds
Low-utility, high-risk	AUC < 0.83 and ECE > 0.06	Unsafe automation	Frequent false triggers or misses	Human-in-the-loop ranked review

The table also clarifies that privacy and non-IID data interact to produce risks that are not visible in global averages. Low-volume clients are primarily impacted through minority-state collapse, while distribution-shift clients show instability and drift. High-utility but high-risk clients require calibration governance rather than architectural changes, whereas low-utility and high-risk clients require conservative deployment with human oversight until additional data, tuning, or personalization closes the reliability gap.

5. Conclusion

The study demonstrates that federated learning is a viable foundation for privacy-preserving adaptive learning analytics in cross-silo educational settings. Across heterogeneous institutions, secure aggregation maintains institutional confidentiality with negligible impact on utility, while update-level differential privacy introduces a predictable privacy-utility tradeoff. The results show that acceptable discrimination can be retained under moderate privacy strength, enabling core analytics functions such as mastery estimation and engagement risk detection without centralizing sensitive learner traces.

The findings also establish calibration as a decisive constraint for deployment in adaptive learning systems. As privacy noise increases, probability reliability deteriorates even when ranking metrics remain competitive, which directly affects threshold-based interventions and automated decision policies. Client-level diagnostics reveal that these effects are uneven across institutions, with low-volume and distribution-shift clients experiencing larger declines and higher decision risk. This supports calibration-aware governance, client-local recalibration, and client-specific thresholding as standard operational controls.

Personalization emerges as a practical mechanism to mitigate non-IID effects and improve equity under privacy constraints. Lightweight client heads increase macro-F1, reduce dispersion across institutions, and partially recover calibration stability, improving the safety and consistency of adaptive interventions. Collectively, the evidence supports a deployment posture that treats privacy parameters, participation scheduling, and personalization as jointly optimized design variables, with client-level monitoring and failure-mode mitigation embedded into the lifecycle of adaptive learning analytics.

6. Declarations

6.1. Author Contributions

Conceptualization: E.Y. and B.G.; Methodology: B.G.; Software: E.Y.; Validation: E.Y. and B.G.; Formal Analysis: E.Y. and B.G.; Investigation: E.Y.; Resources: B.G.; Data Curation: B.G.; Writing Original Draft Preparation: E.Y. and B.G.; Writing Review and Editing: B.G. and E.Y.; Visualization: E.Y.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-Preserving Learning Analytics: Challenges and Techniques," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 68–81, Jan. 2017, doi: 10.1109/TLT.2016.2607747.

[2] P. Prinsloo, S. Slade, and M. Khalil, "The answer is (not only) technological: Considering student data privacy in learning analytics," *British Journal of Educational Technology*, vol. 53, no. 4, pp. 876–893, Jul. 2022, doi: 10.1111/bjet.13216.

[3] C. Fernández-Morante, B. Cebreiro-López, M.-J. Rodríguez-Malmierca, and L. Casal-Otero, "Adaptive Learning Supported by Learning Analytics for Student Teachers' Personalized Training during in-School Practices," *Sustainability*, vol. 14, no. 1, p. 124, Dec. 2021, doi: 10.3390/su14010124.

[4] D. Amo-Filva, D. Fonseca, F. J. García-Peñalvo, M. A. Forment, M. J. Casany Guerrero, and G. Godoy, "Exploring the landscape of learning analytics privacy in fog and edge computing: A systematic literature review," *Computers in Human Behavior*, vol. 158, no. September, p. 108303, Sep. 2024, doi: 10.1016/j.chb.2024.108303.

[5] K. Bonawitz et al., "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas, TX, USA: ACM, Oct. 2017, vol. 2017, no. October, pp. 1175–1191, doi: 10.1145/3133956.3133982.

[6] M. Abadi et al., "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria: ACM, Oct. 2016, vol. 2016, no. October, pp. 308–318, doi: 10.1145/2976749.2978318.

[7] Q. Liu, R. Shakya, M. Khalil, and J. Jovanovic, "Advancing privacy in learning analytics using differential privacy," in *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, Dublin, Ireland: ACM, Mar. 2025, vol. 2025, no. March, pp. 181–191, doi: 10.1145/3706468.3706493.

[8] H. K. Tayyeh and A. S. A. AL-Jumaili, "Balancing Privacy and Performance: A Differential Privacy Approach in Federated Learning," *Computers*, vol. 13, no. 11, p. 277, Oct. 2024, doi: 10.3390/computers13110277.

[9] J. Hu and H. Zhang, "FGS-FL: Enhancing federated learning with differential privacy via flat gradient stream," *Expert Systems with Applications*, vol. 288, no. September, p. 128273, Sep. 2025, doi: 10.1016/j.eswa.2025.128273.

[10] Y. Li, J. Lai, R. Zhang, and M. Sun, "Secure and efficient multi-key aggregation for federated learning," *Information Sciences*, vol. 654, no. January, p. 119830, Jan. 2024, doi: 10.1016/j.ins.2023.119830.

[11] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," *Scientific Data*, vol. 4, no. 1, p. 170171, Nov. 2017, doi: 10.1038/sdata.2017.171.

[12] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 3, Jan. 2023, doi: 10.1186/s41239-022-00371-5.

[13] Z. Huang and C. Cao, "Federated Learning for AI-Assisted Education: Privacy-Preserving and Cross-Platform Collaborative Modeling in Higher Education," in *Proceedings of the 2nd International Conference on Machine Intelligence and Digital Applications*, Ningbo, China: ACM, Apr. 2025, vol. 2025, no. April, pp. 146–151, doi: 10.1145/3744464.3744487.

[14] Q. Liu, R. Shakya, J. Jovanovic, M. Khalil, and J. De La Hoz-Ruiz, "Ensuring privacy through synthetic data generation in education," *British Journal of Educational Technology*, vol. 56, no. 3, pp. 1053–1073, May 2025, doi: 10.1111/bjet.13576.

[15] Z. Li and J. Zhang, "Grouped Federated Learning Algorithm Based on Non-IID Data," in *2023 4th International Conference on Machine Learning and Computer Application*, Hangzhou, China: ACM, Oct. 2023, vol. 2023, no. October, pp. 658–663, doi: 10.1145/3650215.3650331.

[16] P. Kairouz and H. B. McMahan, "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, Jun. 2021, doi: 10.1561/2200000083.

[17] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, Mar. 2019, doi: 10.1145/3298981.

[18] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," *Computers & Security*, vol. 110, no. November, p. 102402, Nov. 2021, doi: 10.1016/j.cose.2021.102402.

[19] J. Zhao et al., "The Federation Strikes Back: A Survey of Federated Learning Privacy Attacks, Defenses, Applications, and Policy Landscape," *ACM Computing Surveys*, vol. 57, no. 9, pp. 1–37, Sep. 2025, doi: 10.1145/3724113.

[20] Y. Gu, Y. Bai, and S. Xu, "CS-MIA: Membership inference attack based on prediction confidence series in federated learning," *Journal of Information Security and Applications*, vol. 67, no. June, p. 103201, Jun. 2022, doi: 10.1016/j.jisa.2022.103201.

[21] A. Hatamizadeh et al., "Do Gradient Inversion Attacks Make Federated Learning Unsafe?," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 2044–2056, Jul. 2023, doi: 10.1109/TMI.2023.3239391.

[22] F. Sabah, Y. Chen, Z. Yang, M. Azam, N. Ahmad, and R. Sarwar, "Model optimization techniques in personalized federated learning: A survey," *Expert Systems with Applications*, vol. 243, no. June, p. 122874, Jun. 2024, doi: 10.1016/j.eswa.2023.122874.

[23] H. Chen, T. Zhu, T. Zhang, W. Zhou, and P. S. Yu, "Privacy and Fairness in Federated Learning: On the Perspective of Tradeoff," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–37, Feb. 2024, doi: 10.1145/3606017.

[24] W. Huang, T. Li, D. Wang, S. Du, J. Zhang, and T. Huang, "Fairness and accuracy in horizontal federated learning," *Information Sciences*, vol. 589, no. April, pp. 170–185, Apr. 2022, doi: 10.1016/j.ins.2021.12.102.

[25] S. K. Banihashem, O. Noroozi, S. Van Ginkel, L. P. Macfadyen, and H. J. A. Biemans, "A systematic review of the role of learning analytics in enhancing feedback practices in higher education," *Educational Research Review*, vol. 37, no. November, p. 100489, Nov. 2022, doi: 10.1016/j.edurev.2022.100489.