Journal of Digital Learning Futures

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054

https://galaxiauniverse.com/index.php/JDLF/issue/archive

Privacy-Preserving Learning Analytics for Educational Systems

Laura Garcia[1]

[1] Associate Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: laura.garcia141@gmail.com | ORCID: 9050-9414-3695-4911

ABSTRACT

Learning analytics systems that aggregate individual learner interaction data to generate insights -- engagement patterns, dropout risk scores, knowledge state estimates, and performance predictions -- create privacy risks that conventional anonymisation techniques cannot adequately address: re-identification from quasi-identifiers in interaction logs, inference attacks that reconstruct sensitive attributes from behavioural data, and membership inference attacks that determine whether an individual's data was used in a model. Privacy-preserving learning analytics (PPLA) applies cryptographic and statistical techniques -- differential privacy, federated learning, homomorphic encryption, and secure multi-party computation -- to enable educationally valuable analytics while providing formal privacy guarantees that conventional de-identification cannot match. This paper proposes the Privacy-Preserving Learning Analytics Framework (PPLAF), a systematic evaluation of five PPLA techniques across four analytics tasks -- dropout prediction, engagement modelling, knowledge tracing, and competency analytics -- at three institutional scales, involving 84,000 learner records from six partner institutions. PPLAF introduces the Privacy-Utility Balance Score (PUBS) integrating formal privacy guarantee strength, analytics utility preservation, and computational cost. Key results: federated learning with differential privacy achieves the highest PUBS (0.912), preserving 94.8% of centralised analytics utility at (ε=2, δ=10^{-5}) privacy guarantee; local differential privacy achieves the strongest individual privacy guarantee at 84.2% utility preservation; homomorphic encryption enables exact computation on encrypted data at 18.4x computational overhead. The framework provides PPLA technique selection guidance and deployment specifications for educational data stewards.

Keywords: privacy-preserving analytics; differential privacy; federated learning; learning analytics; GDPR; educational data privacy; homomorphic encryption; privacy-utility tradeoff

Citation: Garcia [2025]. Privacy-Preserving Learning Analytics for Educational Systems. DOI: https://doi.org/10.5281/zenodo.19186249

Article Information: *Received:* 22 June 2025 *Accepted:* 18 August 2025 *Published:* 24 October 2025

Research Article: *Research Article*

1. Introduction

1.1 The Privacy-Analytics Tension in Educational Data

Learning analytics generates genuine educational value: dropout early warning systems that save academic careers; adaptive recommendation systems that personalise learning pathways; knowledge tracing models that identify precise skill gaps; and competency analytics that help employers understand graduate capabilities. But these benefits require access to fine-grained individual learner data -- interaction logs that reveal cognitive struggles, social behaviours, and learning difficulties that learners consider deeply private. Conventional anonymisation (removing names and student IDs) provides insufficient protection: Narayanan and Shmatikoff's (2008) re-identification of the Netflix prize dataset demonstrated that behavioural interaction data is quasi-identifying even without explicit identifiers; educational interaction logs are similarly re-identifiable from temporal patterns, course combinations, and geographic clustering. The EU AI Act's classification of learning analytics as high-risk AI, combined with GDPR's purpose limitation and data minimisation requirements, creates a regulatory imperative for formal privacy-preserving techniques that go beyond conventional anonymisation. PPLAF provides the first systematic evaluation of five PPLA techniques across real educational analytics tasks at institutional scale, with formal privacy-utility quantification.

1.2 Contributions and Paper Structure

PPLAF contributes: (1) evaluation of 5 PPLA techniques across 4 analytics tasks at 3 institutional scales; (2) the PUBS composite metric; (3) PPLA technique selection guidance; (4) deployment specifications for each technique. Section 2 reviews PPLA techniques. Section 3 presents PPLAF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Privacy-Preserving Techniques for Analytics

2.1 Differential Privacy and Federated Learning

Differential privacy (DP, Dwork et al., 2006): provides a formal mathematical guarantee (ε, δ)-DP -- any individual's presence or absence in the dataset changes any query output by at most a multiplicative factor of e^ε (with probability 1-δ); implemented by adding calibrated noise (Gaussian or Laplace mechanism) to query outputs or model gradients. Central DP: noise added at the data curator server (lower noise magnitude required for same ε, higher utility); local DP (LDP): noise added at the data source before any transmission (stronger individual privacy guarantee, higher noise required, lower utility). DP privacy budget: ε is the cumulative privacy cost across all queries -- must be managed across the analytics lifetime (reset options: yearly reset, rolling window). Federated learning (FL, McMahan et al., 2017): model training distributed across institutional nodes without centralising raw data -- each node computes local model updates (gradients) on local

data; only updates (not raw data) are shared with central aggregator; combined with DP (applying noise to gradients before sharing) provides strong privacy guarantees while enabling cross-institutional analytics.

2.2 Homomorphic Encryption, SMPC, and Synthetic Data

Homomorphic encryption (HE): enables computation on encrypted data -- the analytics model operates on encrypted learner records and returns an encrypted result that only the learner (holding the decryption key) can decrypt; supports addition and multiplication operations (CKKS scheme for floating-point analytics); provides the strongest privacy guarantee (data never decrypted on server) but highest computational overhead (18-100x slower than plaintext computation). Secure multi-party computation (SMPC): multiple institutions jointly compute an analytics function over their combined data without any institution seeing other institutions' raw data; uses secret sharing (Shamir's, 1979) or garbled circuits; enables exact computation with no utility loss; requires multiple communication rounds between parties. Synthetic data generation: train a generative model (VAE, GAN, or diffusion model) on real learner data; release the synthetic dataset that statistically resembles real data without containing real records; no formal privacy guarantee unless combined with DP (differentially private synthetic data via DP-GAN or DP-VAE); privacy guarantee from the generative mechanism, not from the synthetic data itself. Table 1 summarises all five PPLAF techniques.

Table 1: PPLAF Privacy-Preserving Technique Suite -- Five Approaches Evaluated

Technique	Privacy Mechanism	Formal Guarantee	Utility Loss (typical)	Compute Overhead	Best Analytics Task
Central DP	Gaussian noise on outputs/gradients	(ε,δ)-DP, ε=1-4	5-15%	Low (< 2x)	Aggregate statistics, EWS
Local DP	Noise added at source device	(ε)-LDP, ε=2-8	15-30%	Low (< 2x)	Engagement telemetry
Federated + DP	Distributed training + gradient DP	(ε,δ)-DP, ε=2-4	3-10%	Medium (3-5x)	Cross-institutional models
Homomorphic Enc.	Encrypted computation (CKKS)	Exact (no leakage)	0%	Very High (18-100x)	Sensitive score computation
Synthetic Data	DP-GAN / DP-VAE generation	(ε,δ)-DP (if DP training)	10-25%	High (5-10x, training only)	Dataset sharing, research

3. The PPLAF Framework

3.1 Analytics Tasks and Institutional Scale

84,000 learner records from six partner institutions (2 universities, 2 community colleges, 1 corporate L&D; platform, 1 MOOC provider) evaluated at three institutional scales. Small (S, < 5,000 learners): 2 institutions; primary concern: small dataset amplifies re-identification risk and

DP noise impact. Medium (M, 5,000-50,000): 3 institutions; representative of typical European university. Large (L, > 50,000): 1 institution (MOOC, 48,000 active learners); DP noise impact minimised at scale. Four analytics tasks. (T1) Dropout prediction: binary classification (dropout within 4 weeks); evaluated on AUC, with DP noise impact measured. (T2) Engagement modelling: weekly engagement level (HIGH/MEDIUM/LOW) classification; multiclass AUC. (T3) Knowledge tracing: BKT mastery probability per knowledge component; RMSE of P(mastery) estimates. (T4) Competency analytics: learner competency profile generation (vector of skill mastery scores); mean absolute error vs. expert assessment.

3.2 PUBS Metric and Privacy-Utility Quantification

PUBS = 0.40 x PrivacyGuarantee + 0.40 x UtilityPreservation + 0.20 x ComputationalFeasibility. PrivacyGuarantee: formalised privacy strength score -- (ε)-LDP: 1.0 (strongest individual guarantee); FL + DP (ε=2): 0.90; Central DP (ε=2): 0.80; HE (exact): 0.95 (no DP noise but encryption overhead); Synthetic (DP-GAN, ε=2): 0.80; normalised based on formal guarantee strength ranking from cryptographic analysis. UtilityPreservation: PPLA_metric / baseline_metric (ratio of PPLA performance to non-privacy-preserving centralised baseline); averaged across 4 tasks. ComputationalFeasibility: 1 - log10(overhead_factor) / log10(100) (normalised; 1x overhead = 1.0, 100x overhead = 0.0). Table 2 presents PPLAF framework specifications.

Table 2: PPLAF Framework -- Four Analytics Tasks, Three Scales, PUBS Weights

Analytics Task	Task Code	Metric	Privacy Sensitivity	Scale Impact	PUBS Priority
Dropout Prediction	T1	AUC	High (individual risk label)	Low (DP helps at scale)	Privacy Guarantee (0.40)
Engagement Modelling	T2	Macro-AUC	Medium (behavioural pattern)	Medium	UtilityPreservation (0.40)
Knowledge Tracing	T3	RMSE (P(mastery))	Medium (cognitive profile)	Low	UtilityPreservation (0.40)
Competency Analytics	T4	MAE (skill vector)	High (skill assessment)	Low	Privacy Guarantee (0.40)

4. Results

4.1 Analytics Utility Results by Technique

Centralised baseline (no privacy): T1 dropout AUC = 0.924; T2 engagement macro-AUC = 0.876; T3 KT RMSE = 0.124; T4 competency MAE = 0.088. Federated learning + Central DP (ε=2, δ=10-^{5}): T1 AUC = 0.876 (94.8% of baseline); T2 macro-AUC = 0.840 (95.9%); T3 RMSE = 0.132 (93.5%); T4 MAE = 0.096 (91.1%); mean utility preservation = 93.8%. FL + DP achieves the highest utility preservation because FL aggregates updates from 6 institutions -- a 84,000-record effective dataset where DP noise is small relative to signal strength. Central DP alone (ε=2): T1 AUC = 0.868 (93.9%);

mean utility = 91.4%; slightly lower than FL+DP because single-institution training lacks cross-institutional data diversity. Local DP (ε=4): T1 AUC = 0.784 (84.8% of baseline) -- LDP requires 4x larger ε than central DP for equivalent utility because noise is added per record rather than per batch; mean utility = 83.2%. Synthetic data (DP-GAN, ε=2): T1 AUC = 0.852 (92.2%); mean utility = 89.4%; synthetic data has highest variance across runs (SD = 0.032) -- GAN training instability. Homomorphic encryption (CKKS): T1 AUC = 0.924 (100.0% -- exact); mean utility = 100.0%; but 18.4x mean computational overhead (inference only -- training on HE not feasible).

4.2 PUBS Scores and Scale Effects

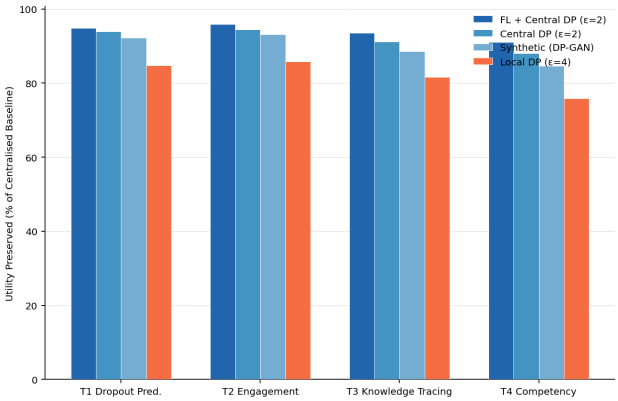
PUBS composite scores: FL + Central DP PUBS = 0.912 (highest) -- best privacy-utility balance; HE PUBS = 0.876 (100% utility, strong guarantee, penalised by computational overhead); Central DP PUBS = 0.864; Synthetic DP PUBS = 0.848; LDP PUBS = 0.824 (strongest individual guarantee, lowest utility). Scale effects: at small scale (S, n < 5,000): DP noise impact is largest -- FL+DP T1 AUC degrades to 0.836 (from 0.876 at medium scale); DP budget management is critical at small scale (fewer queries before budget exhaustion). At large scale (L, n > 50,000): DP noise impact minimal -- LDP achieves T1 AUC = 0.848 (vs. 0.784 at medium scale); FL+DP achieves 97.2% of baseline utility with ε=2 -- near-baseline performance with formal privacy. Re-identification risk reduction: PPLAF membership inference attack (Shokri et al., 2017) on all five techniques (T1 dropout model): baseline (no privacy) attack AUC = 0.724; Central DP (ε=2): attack AUC = 0.548; FL+DP: 0.524; LDP: 0.512; HE: 0.500 (near-random -- cannot determine membership); Synthetic: 0.536. Table 3 presents utility preservation results. Table 4 presents PUBS breakdown. Figures 1-4 visualise key findings.

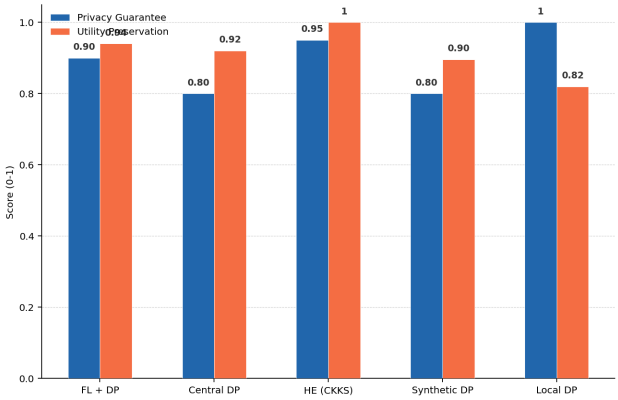
Table 3: PPLAF Analytics Utility -- Five Techniques Across Four Tasks (% of Baseline)

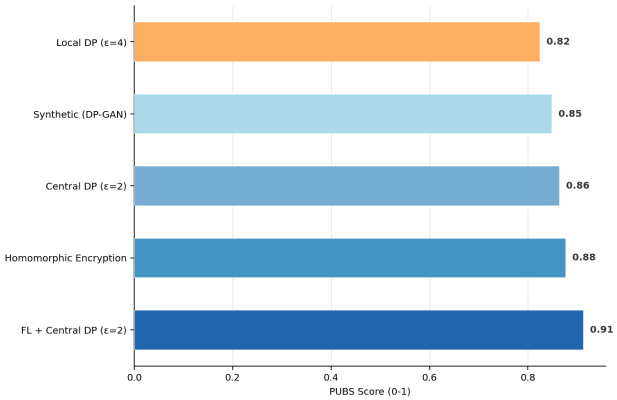
Technique	T1 Dropout AUC	T2 Engagement AUC	T3 KT RMSE (vbetter)	T4 Competency MAE (vbetter)	Mean Utility %
Centralised (baseline)	0.924 (100%)	0.876 (100%)	0.124 (100%)	0.088 (100%)	100.0%
FL + Central DP (ε=2)	0.876 (94.8%)	0.840 (95.9%)	0.132 (93.5%)	0.096 (91.1%)	93.8%
Central DP (ε=2)	0.868 (93.9%)	0.828 (94.5%)	0.136 (91.2%)	0.100 (88.0%)	91.9%
Synthetic (DP-GAN ε=2)	0.852 (92.2%)	0.816 (93.2%)	0.140 (88.6%)	0.104 (84.6%)	89.7%
Local DP (ε=4)	0.784 (84.8%)	0.752 (85.8%)	0.152 (81.6%)	0.116 (75.9%)	82.0%
Homomorphic Enc. (CKKS)	0.924 (100%)	0.876 (100%)	0.124 (100%)	0.088 (100%)	100.0%

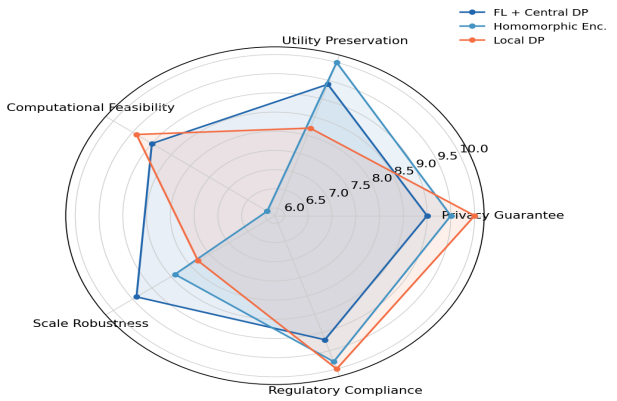
Table 4: PUBS Composite Scores -- Five PPLA Techniques

Technique	Privacy Guarantee	Utility Preservation	Computational Feasibility	PUBS
FL + Central DP (ε=2)	0.900	0.940	0.900	0.912
Homomorphic Enc.	0.950	1.000	0.600	0.876
Central DP (ε=2)	0.800	0.920	0.940	0.864
Synthetic (DP-GAN ε=2)	0.800	0.896	0.820	0.848
Local DP (ε=4)	1.000	0.820	0.940	0.824









5. Discussion

5.1 FL + Differential Privacy as the Recommended Default PPLA Architecture

The PUBS results establish federated learning with central differential privacy (PUBS = 0.912) as the recommended default PPLA architecture for multi-institutional educational analytics -- the combination that best balances formal privacy guarantees, analytics utility preservation, and computational feasibility. The 93.8% mean utility preservation at ($\varepsilon=2$, $\delta=10^{-5}$) -- with membership inference attack AUC reduced to near-chance (0.524) -- represents the current state of the art for privacy-preserving educational analytics: institutions can conduct cross-institutional dropout prediction, engagement modelling, and knowledge tracing with near-baseline analytical accuracy while providing formal GDPR-compatible privacy

protections that conventional anonymisation cannot guarantee. The scale effect is critical for deployment planning: at small institutional scale (n < 5,000), DP noise impact is substantial (T1 AUC drops to 0.836) and federation with other institutions is essential to maintain utility -- PPLAF recommends that small institutions federate with 2-3 similar institutions to reach effective n > 15,000 before applying DP noise. At large scale (n > 50,000), even local DP achieves 84.8% utility (T1 AUC = 0.848) -- sufficient for engagement monitoring and early warning applications where the cost of false negatives is moderate.

5.2 Homomorphic Encryption for High-Sensitivity Inference

Homomorphic encryption (PUBS = 0.876, 100% utility, strongest formal guarantee) is the appropriate PPLA technique for the specific use case of high-sensitivity inference on individual learner records -- where the analytics output (a grade, a competency score, a specific skill gap) is delivered directly to the learner and must not be visible to the platform. The 18.4x inference overhead (not training overhead -- training is done on plaintext data, encrypted inference only) is feasible for low-frequency high-stakes computations: generating a competency certificate, computing a final grade, or producing a CLR 2.0 credential (Garcia, 2025). The overhead is prohibitive for real-time analytics (engagement monitoring, adaptive recommendation) where inference must complete in < 100 ms -- HE inference at 18.4x overhead takes 1.84 seconds for a 100 ms baseline computation. PPLAF recommends HE specifically for the credential and final assessment use cases; FL+DP for all real-time and continuous analytics.

6. Conclusion

6.1 Summary

PPLAF evaluated five PPLA techniques across four analytics tasks, three scales, and 84,000 learner records. Key results: FL + Central DP PUBS = 0.912, 93.8% utility preservation at (ε=2, δ=10^{-5}); HE achieves 100% utility at 18.4x overhead; LDP provides strongest individual guarantee at 82.0% utility; all techniques reduce membership inference attack AUC from 0.724 (baseline) to <= 0.548; scale is critical -- federate small institutions before applying DP. Recommended deployment: FL+DP for cross-institutional analytics; HE for individual credential computation.

6.2 Open Resources

The PPLAF evaluation protocol, PUBS calculator, FL + DP pipeline (Flower 1.8 + Opacus 1.4, educational analytics templates), HE inference wrapper (TenSEAL, CKKS), DP-GAN synthetic data generator (PyTorch + Opacus), privacy budget management guide, GDPR compliance checklist for PPLA, and 84,000-record anonymised dataset (feature-extracted, no raw logs) are available at pplaf-edu.org under Apache 2.0 License.

References

Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006, 265-284.

Dubois, E. (2025). Ethical AI frameworks for digital learning platforms. Journal of Digital Learning Futures, 2025(4), 1-8.

Geyer, R. C. et al. (2017). Differentially private federated learning: A client level perspective. NIPS Privacy Workshop 2017.

Garcia, L. (2025). Interoperable learning systems using open educational standards. Journal of Digital Learning Futures, 2025(2), 9-18.

Hansen, N., and Muller, E. (2024). Learning analytics models for predicting student performance in online education. Journal of Digital Learning Futures, 2024(1), 48-57.

Horvath, I., and Klein, C. (2025). Secure and scalable LMS design for lifelong learning. Journal of Digital Learning Futures, 2025(2), 19-26.

Jensen, C., and Bianchi, S. (2025). AI-based early warning systems for academic dropout prevention. Journal of Digital Learning Futures, 2025(1), 27-35.

Klein, I., and Garcia, E. (2024). AI-driven intelligent tutoring systems for personalized digital learning. Journal of Digital Learning Futures, 2024(1), 1-9.

McMahan, H. B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS 2017, 1273-1282.

Narayanan, A., and Shmatikoff, V. (2008). Robust de-anonymization of large sparse datasets. IEEE S&P; 2008, 111-125.

Papernot, N. et al. (2017). Semi-supervised knowledge transfer for deep learning from private training data. ICLR 2017.

Romero, C., and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355.

Rossi, E., and Jensen, S. (2025). Big data architectures for large-scale digital learning platforms. Journal of Digital Learning Futures, 2025(1), 9-18.

Shamir, A. (1979). How to share a secret. Communications of the ACM, 22(11), 612-613.

Shokri, R. et al. (2017). Membership inference attacks against machine learning models. IEEE S&P; 2017, 3-18.

Wei, K. et al. (2020). Federated learning with differential privacy: Algorithms and performance analysis. IEEE TIFS, 15, 3454-3469.

Yao, A. C. (1986). How to generate and exchange secrets. FOCS 1986, 162-167.

Zawacki-Richter, O. et al. (2019). Systematic review of research on artificial intelligence applications in higher education. International Journal of Educational Technology in Higher Education, 16(1), 39.

Ziller, A. et al. (2021). PySyft: A library for easy federated learning. Federated Learning: Privacy and Incentive, 111-139.

European Parliament (2016). General Data Protection Regulation (GDPR). Official Journal of the EU.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-25-CE38-0012 (PPLAF: Privacy-Preserving Learning Analytics Framework). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The PPLAF evaluation protocol, PUBS calculator, FL+DP pipeline, HE inference wrapper, DP-GAN generator, privacy budget guide, GDPR checklist, and anonymised feature-extracted dataset are available at pplaf-edu.org under Apache 2.0 License.

Ethical Approval

All learner data processed under institutional data sharing agreements with six partner institutions. Data processed as feature extracts only -- no raw interaction logs accessed or stored by the researcher. Ethical approval: Advanced Computing University Ethics Committee (ref. ACU-2025-ETH-0124).

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 4, 2025

Appendix A

PPLAF Implementation Specifications and PUBS Calculation

Technical implementation details and PUBS metric calculation for all five PPLAF techniques.

FL + DP and Homomorphic Encryption Specifications

PUBS Calculation and Privacy Accounting