

Privacy-Preserving Artificial Intelligence in Digital Learning: A Systematic Review of Federated Learning, Differential Privacy, Security, and Educational Utility in Higher Education

ADESEGUN NURUDEEN OSIJIRIN

adesegunosijirin@fuahse.edu.ng

Federal University of Allied Health Sciences, Enugu

Oliver Okechukwu

Enugu State University of Science and Technology

Leonard C. Anigbo

Enugu State University of Science and Technology

Systematic Review

Keywords: privacy-preserving artificial intelligence, federated learning, differential privacy, learning analytics, higher education, educational data mining, synthetic data

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10922594/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence (AI) and learning analytics increasingly depend on sensitive learner records, yet centralised data pipelines intensify privacy, security and institutional-governance risks. This systematic review synthesised empirical evidence on privacy-preserving AI in digital learning, with emphasis on higher education. OpenAlex and Crossref were searched from January 2018 to 3 September 2026 using 11 query families covering federated learning, differential privacy, synthetic educational data, secure computation, homomorphic encryption and local AI. Targeted searches of publisher indexes and backward/forward citation checking supplemented retrieval. Following PRISMA 2020, 3,850 records were identified, 2,923 unique records were screened and 33 studies met the eligibility criteria. Because tasks, datasets and privacy measures were heterogeneous, findings were synthesised narratively using a structured privacy-security-utility framework. Federated learning was the primary mechanism in 15 studies, differential privacy in five, synthetic-data generation in seven, cryptographic/secure-edge computation in two, and data-minimising or locally hosted AI in four. Most studies reported utility close to centralised or non-private baselines; however, only a minority evaluated privacy through explicit budgets, re-identification or membership-inference tests. Security threats such as gradient leakage, poisoning, malicious aggregation and collusion were rarely examined. Public benchmarks—especially OULAD—were repeatedly reused, while live cross-institution deployment, cost, governance, learner consent and causal learning outcomes were largely absent. Privacy-preserving AI is technically promising but not yet institutionally validated. Future studies should combine explicit threat models, differential-privacy accounting, secure aggregation, non-IID and subgroup evaluation, external validation, and educational-impact assessment in real multi-university settings.

1. Introduction

Digital learning platforms now record fine-grained traces of participation, assessment, writing, collaboration, navigation and help-seeking. These records can support early-warning systems, adaptive feedback, course recommendation and institutional planning, but they can also expose academic performance, identity, disability, language, behaviour and inferred psychological states. Learning analytics therefore presents a structural tension: the same granularity that may improve prediction can increase surveillance, re-identification and discriminatory use (Slade & Prinsloo, 2013; Pardo & Siemens, 2014; Ifenthaler & Schumacher, 2016).

The regulatory environment raises the stakes. The General Data Protection Regulation establishes principles of purpose limitation, data minimisation, transparency and data-subject rights, while the European Union Artificial Intelligence Act adds risk-management, documentation and human-oversight obligations for many educational AI uses (European Parliament & Council of the European Union, 2016, 2024). UNESCO (2023) similarly emphasises human agency, inclusion and data protection in educational AI. Although these instruments differ in jurisdiction and legal force, they converge on a practical requirement: universities must justify not only what an AI system predicts, but also what data it uses,

where those data travel, who can infer information from model outputs and whether the resulting intervention benefits learners.

Privacy-enhancing technologies offer several possible responses. Federated learning trains a shared model while keeping raw records at participating clients or institutions (McMahan et al., 2017). Differential privacy limits how strongly any single record can affect an output through calibrated randomisation (Dwork & Roth, 2014). Secure aggregation prevents a coordinating server from inspecting individual model updates (Bonawitz et al., 2017). Homomorphic encryption and secure multiparty computation allow selected operations over encrypted or distributed inputs. Synthetic data and local or on-device models can further reduce disclosure by limiting access to identifiable records.

These techniques are complementary rather than interchangeable. Federated learning reduces raw-data movement but does not, by itself, stop gradient inversion, membership inference, poisoning or leakage from model updates (Kairouz et al., 2021). Differential privacy offers a quantifiable guarantee, but aggressive noise may degrade utility, especially for small classes or minority groups. Synthetic data can enable sharing yet reproduce bias or leak training examples. Cryptographic methods can strengthen confidentiality while adding computational and operational burdens. The central evaluative problem is therefore not whether privacy is present, but what privacy guarantee is provided, against which adversary, at what utility and institutional cost.

Earlier learning-analytics scholarship mapped ethical and privacy concerns, stakeholder expectations and governance principles. A systematic review by Liu and Khalil (2023) identified a shortage of applied evidence and called for practical evaluation of privacy solutions. Since then, empirical research has expanded quickly, spanning federated at-risk prediction, differentially private knowledge tracing, private synthetic data, encrypted teaching-evaluation analysis and locally hosted language models. This growth creates a need for an updated evidence synthesis that evaluates privacy assurance and educational usefulness together, rather than treating technical performance or policy compliance in isolation.

1.1 Review objective and questions

This review examines whether privacy-preserving AI for digital learning has progressed from promising architectures to credible, educationally useful and deployable evidence. It addresses five questions:

1. Which privacy-preserving technologies and educational AI tasks have been empirically evaluated?
2. What privacy guarantees, threat models and security controls are reported?
3. How does privacy protection affect predictive, analytical or pedagogical utility?
4. To what extent are fairness, interpretability, reproducibility and institutional deployment evaluated?
5. What evidence and governance gaps should guide future higher-education research?

1.2 Conceptual boundaries

In this review, privacy refers to limiting information disclosure about learners or institutions; security refers to protecting the confidentiality, integrity and availability of data and model processes; and utility refers to the usefulness of a protected system for its stated analytical or educational task. Privacy claims were classified as: (a) formal, when supported by a mathematical guarantee such as an explicit differential-privacy budget or cryptographic construction; (b) empirical, when supported by an attack, re-identification or disclosure-risk test; or (c) architectural, when inferred mainly from decentralisation, masking, deletion or local processing. This distinction is essential because architectural data locality can reduce exposure without establishing resistance to inference from shared parameters or outputs.

2. Methods

2.1 Design and reporting

A systematic review with narrative synthesis was conducted and reported in accordance with PRISMA 2020 and PRISMA-S principles (Page et al., 2021; Rethlefsen et al., 2021). The protocol specified the review questions, eligibility criteria, search concepts, extraction fields and synthesis framework before final study selection. The protocol was not prospectively registered; the full search logic and appraisal domains are reproduced in Appendix A. Meta-analysis was not planned because included studies differed substantially in educational tasks, unit of federation, privacy mechanisms, outcome definitions, baselines and datasets. Narrative synthesis followed the logic of the SWiM reporting guideline (Campbell et al., 2020).

2.2 Information sources and search strategy

OpenAlex and Crossref were searched programmatically for peer-reviewed journal and conference literature published between 1 January 2018 and 3 September 2026. Eleven overlapping query families combined privacy-preserving technology terms with educational context and analytical task terms. Targeted searches of ACM Digital Library, IEEE Xplore, ERIC, SpringerLink, ScienceDirect and publisher pages were used to verify metadata, locate full reports and identify records not reliably indexed in the two discovery databases. Reference lists and citation links of eligible studies and the review by Liu and Khalil (2023) were also checked. Searches were restricted to English-language records. Up to 200 ranked OpenAlex results and 150 Crossref results were retained per query family; this retrieval cap is considered in the limitations.

Table 1
Core search concepts and representative terms

Concept	Representative search terms
Privacy technology	"federated learning"; "differential privacy"; "privacy-preserving machine learning"; "secure multiparty computation"; "homomorphic encryption"; "synthetic student data"; "locally hosted"; "on-device"
Educational context	education; higher education; digital learning; e-learning; learning analytics; educational data mining; classroom; course; learner; student
AI/analytics task	student performance; dropout; at-risk prediction; recommendation; knowledge tracing; grading; engagement; teaching evaluation; learning analytics

2.3 Eligibility criteria

Studies were included when they: (1) reported a peer-reviewed full journal article or full conference paper; (2) implemented or empirically evaluated an AI, machine-learning or learning-analytics system using educational data; (3) applied a concrete privacy-preserving mechanism to learner, teaching, assessment or institutional data; (4) reported at least one utility or system-performance outcome; and (5) concerned higher education, online courses, learning-management systems, health-professions education or a digital-learning dataset with plausible postsecondary relevance. Studies spanning more than one educational level were retained when higher-education or platform-level evidence could be separated.

Reviews, editorials, protocols, theses, abstracts, preprints without a peer-reviewed version, blockchain credentialing without AI analytics, generic cybersecurity systems, purely legal analyses, healthcare-only federated learning and papers that merely described privacy concerns were excluded. Multiple versions of the same study were consolidated in favour of the peer-reviewed record.

2.4 Study selection

Automated normalisation removed duplicates using DOI and normalised title. A broad rule-based relevance pass identified records containing both educational and privacy-technology concepts. Records were then assessed against the predefined criteria, with uncertain records advanced to the next stage. The workflow used automated deduplication and keyword triage; it did not provide independent dual-reviewer screening. Of 3,850 records identified, 927 duplicates were removed, 2,923 titles and abstracts were screened, 117 full reports were assessed and 33 studies were included (Fig. 1).

2.5 Data extraction and evidence appraisal

A structured extraction matrix captured publication year, setting, data source, educational task, privacy mechanism, federation unit, model family, comparator, privacy evidence, utility outcome, fairness or interpretability analysis, deployment status, reproducibility resources and stated limitations. Findings were checked against DOI metadata and available publisher or author full texts. No conventional clinical risk-of-bias tool fits a corpus dominated by computational experiments. Consequently, evidence

readiness was appraised descriptively across seven domains: data provenance and sample transparency; comparator adequacy; privacy specification; empirical threat evaluation; robustness or external validation; fairness/interpretability; and reproducibility. No study was excluded solely because of weak reporting.

2.6 Synthesis

Studies were grouped by their primary privacy mechanism: federated learning; differential privacy; private synthetic data; cryptographic or secure-edge computation; and data-minimising/local AI. Overlapping controls were recorded separately. Utility was synthesised directionally relative to centralised or non-private baselines. Privacy strength was interpreted according to the formal-empirical-architectural classification in Section 1.2. Counts describe the included corpus and should not be interpreted as effect estimates.

3. Results

3.1 Study selection and growth of the field

Thirty-three studies published between 2018 and September 2026 met the criteria. Publication increased sharply: one study appeared in 2018, two in 2021, two in 2022, four in 2023, three in 2024, ten in 2025 and eleven by early September 2026. Thus, 21 of 33 studies (63.6%) were published in 2025–2026, indicating a rapidly changing evidence base rather than a mature implementation literature.

3.2 Study characteristics

The corpus covered dropout and academic-risk prediction, grade and outcome prediction, recommendation and personalisation, engagement and activity monitoring, teaching-evaluation analysis, knowledge tracing, peer-link prediction, synthetic-data sharing and automated assessment. The most common design was an offline computational experiment using a public dataset or institution-specific historical records partitioned into simulated clients. OULAD (Kuzilek et al., 2017) was repeatedly reused for dropout, risk, synthetic-data and federated-learning experiments. Few studies involved authentic inter-institution data exchange, and none reported sustained production deployment across multiple universities.

Table 2
Included studies published from 2018 to 2024

Study	Context/data	Primary mechanism	Main evidence	Principal limitation
Feng et al. (2018)	Learner-behaviour recommendation	Laplace-noise differential privacy	Recommendation accuracy reportedly maintained while perturbing similarity scores.	No deployment or adversarial test.
Preuveneers et al. (2021)	Multimodal classroom engagement	Browser/edge analytics plus secure multiparty computation	Demonstrated scalable aggregation without centralising raw multimodal data.	Limited field scale and threat testing.
Rodríguez-García et al. (2021)	Cloud e-learning data mashups	Linked data plus k-anonymisation	Retained statistical utility and reduced re-identification in empirical tests.	No formal DP budget; limited ML evaluation.
Flanagan et al. (2022)	Private educational dataset and modelling challenge	Synthetic data sharing	Some models trained on synthetic data transferred poorly to the private holdout.	Shows that fidelity does not ensure transfer.
Chu et al. (2022)	Three online-course datasets	Personalised attention-based federated learning	Improved outcome prediction across demographic subgroups relative to baselines.	No formal privacy accounting.
Qin et al. (2023)	Multimedia-course recommendation	Federated recommender	Supported recommendation without pooling learner records.	Privacy inferred mainly from data locality.
Li et al. (2023)	Teamwork video in simulation-based learning	Computer vision, face masking and deletion controls	Recovered useful socio-spatial features with low-cost hardware.	Privacy-by-design, not formal PPML.
Fachola et al. (2023)	Dropout prediction in learning analytics	Federated learning	Federated models achieved performance comparable to centralised models.	Simulated federation; limited security tests.
Mistry et al. (2023)	> 4,000 e-learning screenshots	Federated InceptionV3	Reported 99.75% test accuracy for on-screen activity classification.	Curated data; privacy equated with FL.
van Haastrecht et al. (2024)	Three datasets and prediction use cases	Local vs federated vs central learning	FL was consistently comparable to central learning; local models	No production cross-

Study	Context/data	Primary mechanism	Main evidence	Principal limitation
			remained competitive up to 20 clients.	institution trial.
Zhan et al. (2024)	Large-scale educational dataset	Multiple privacy-preserving transformations	Quantified utility loss after privacy protection and showed workable trade-offs.	Limited attack coverage.
Liu et al. (2024)	Three learning-analytics datasets	Three synthetic-data generators	Evaluated resemblance, downstream utility and privacy; optimal method depended on use case.	Static tabular focus.

Table 3
Included studies published from 2025 to September 2026

Study	Context/data	Primary mechanism	Main evidence	Principal limitation
Liu et al. (2025a)	OULAD-based learning-analytics scenario	Differential-privacy framework	Demonstrated workable privacy-utility settings, including strong privacy budgets below 1.	Single main benchmark.
Liu et al. (2025b)	Multiple learning-analytics datasets	Synthetic generation plus fairness pre-processing	DECAF offered the best privacy-fairness balance but reduced predictive utility.	No field deployment.
Chen & Qi (2025)	Programming-course records; 10 simulated clients	Entropy-adaptive DP-FedAvg	Accuracy 92.7%; approached centralised performance under reported privacy budgets.	Single institution and simulated clients.
Zhao et al. (2025)	Real educational grade-prediction data	DP-TabNet; epsilon = 0.7	Accuracy 80% versus 84% without privacy; retained feature importance.	Narrow task and dataset scope.
Iloh et al. (2025)	OULAD	Gaussian Copula and CTGAN synthetic data	CTGAN retained 93.1% of dropout-prediction utility with membership-inference testing.	One benchmark; static records.
Kesgin (2025)	MOOC, tutoring and LMS datasets	Diffusion synthesis, DP and adversarial debiasing	Competitive RMSE (0.402) and lower disparities; membership-inference AUC 0.705.	Privacy protection was moderate, not absolute.
Hridi et al. (2025)	Peer networks in online classrooms	Federated link prediction	Enabled cross-classroom link prediction without pooling interaction graphs.	Formal leakage testing absent.
Yoneda et al. (2025)	1,136 students, 12 courses; five holdout courses	FL plus differential features	Comparable central performance and useful early at-risk detection on unseen courses.	Single institutional ecosystem.
Kabir & Lowd (2025)	Knowledge-tracing benchmarks	Differential privacy	Mapped when DP remains practical and showed strong sensitivity to budget and task conditions.	Short conference report.

Study	Context/data	Primary mechanism	Main evidence	Principal limitation
Zhang et al. (2025)	Teaching-evaluation sentiment datasets	CKKS homomorphic encryption plus ALBERT	Competitive classification with a modest computational penalty.	No institutional deployment or user study.
Villegas-Ch et al. (2026)	Heterogeneous institutional partitions	Federated temporal modelling plus explainability	Discrimination remained strong, but calibration degraded under distribution shift.	Controlled rather than live federation.
Jodayree et al. (2026)	Three datasets; primary cohort n = 14,003	FL, client DP and verifiable training log	FL + DP achieved 98.0% accuracy versus 99.14% centralised; explanation patterns were broadly preserved.	Simulated schools and labels.
Chen et al. (2026)	OULAD	CNN-based FL plus adaptive-noise DP	Accuracy exceeded 0.9516 and F1 reached 0.9522 under the tested setting.	One benchmark and unusually high performance.
Agal (2026)	SynEdu-HEDL: 20,000 records, 85 features	GAN/temporal synthesis plus DP	Membership-inference AUC 0.512; downstream models were within 1–5% of real-data utility.	Article in press; source-data generalisability.
Iloh et al. (2026)	OULAD and ASSISTments	Gaussian Copula, CTGAN and TabDDPM	TabDDPM AUC 0.962 on OULAD; worst membership-inference AUC < = 0.527.	Static benchmarks; no fairness analysis.
Rejeleene et al. (2026)	10 medical-student reflections	Locally hosted open-weight LLM	Consistent rubric scoring; faculty rated feedback 5.48/7.	Very small feasibility study; de-identified data.
Saur & Toth (2026)	Synthetic short-answer grading corpus	Fine-tuned local small language models	In-domain macro-F1 up to 0.96 versus 0.90 for GPT-4o; weaker cross-domain transfer.	Synthetic corpus and no learner-outcome test.
Zhou & Yuan (2026)	Eight simulated institutions; learner corpora	Adaptive DP-FL, secure aggregation and personalised layers	Attack success approached chance; recovered roughly three-quarters of the private-to-central utility gap.	Article in press; simulated federation.

Study	Context/data	Primary mechanism	Main evidence	Principal limitation
Kovor et al. (2026)	Health-professions at-risk prediction	FedAvg and FedProx	Multi-scenario comparisons supported competitive federated prediction.	Privacy mainly architectural; no live federation.
Liao (2026)	Online-learning adaptability under non-IID data	Adaptive FL, noise and Top-k compression	Reduced communication by nearly 60% while retaining competitive performance.	Privacy parameterisation under-reported.
Ji et al. (2026)	EdNet, ASSISTments and OULAD	Federated LLM adapters, secure masking and Rényi DP	AUC gains of 4.2–7.1%, up to 28% lower calibration error and fivefold communication reduction.	Public benchmarks and simulated institutions.

3.3 Privacy-preserving mechanisms

Federated learning was the primary mechanism in 15 studies. Across recommendation, dropout, grade, peer-link and outcome prediction, federated models usually approached centralised performance (Qin et al., 2023; Fachola et al., 2023; van Haastrecht et al., 2024; Hridi et al., 2025; Yoneda et al., 2025; Kovor et al., 2026). Personalisation, adaptive aggregation and differential features improved performance under heterogeneous data (Chu et al., 2022; Chen & Qi, 2025; Liao, 2026). More recent work combined federated learning with explainability, differential privacy, secure aggregation or parameter-efficient language-model adaptation (Jodayree et al., 2026; Villegas-Ch et al., 2026; Ji et al., 2026; Zhou & Yuan, 2026).

Differential privacy was the primary mechanism in five studies and appeared as an additional control in several federated or synthetic-data studies. Early recommendation work used Laplace perturbation (Feng et al., 2018), while knowledge-tracing experiments demonstrated strong sensitivity to the selected privacy budget and task conditions (Kabir & Lowd, 2025). Explicit budgets made the privacy-utility trade-off inspectable in studies such as Liu et al. (2025a), Chen and Qi (2025), Zhao et al. (2025), and Ji et al. (2026). However, accounting conventions, neighbouring-dataset definitions, clipping rules and unit of privacy varied, reducing comparability. Some studies described noise-based privacy without reporting enough information to reproduce the guarantee.

Seven studies primarily examined synthetic educational data. The evidence was mixed but informative. Flanagan et al. (2022) showed that models trained on synthetic records could perform significantly worse on private validation data, whereas more recent CTGAN, diffusion and hybrid generators retained high downstream utility and near-chance membership-inference performance in selected settings (Iloh et al., 2025, 2026; Agal, 2026). FairSYN-Edu reduced demographic disparities but still produced a membership-inference AUC of 0.705, underscoring that synthetic does not mean anonymous (Kesgin,

2025). Liu et al. (2024, 2025b) likewise showed that privacy, utility and fairness can conflict and that generator choice should be task-specific.

Cryptographic and secure-edge approaches were less common. Preuveneers et al. (2021) combined browser/edge processing with secure multiparty computation for classroom engagement analysis, while Zhang et al. (2025) used approximate homomorphic encryption for teaching-evaluation sentiment classification. Four studies relied mainly on minimisation, masking, deletion or local hosting. These approaches can materially reduce disclosure—through linked-data anonymisation, for example, or local handling of video and free text—but generally lacked formal privacy accounting (Rodríguez-García et al., 2021; Li et al., 2023; Rejeleene et al., 2026; Saur & Toth, 2026).

3.4 Privacy and security evidence

Nineteen studies provided either an explicit formal parameter, a cryptographic construction or an empirical disclosure-risk test. Eleven reported an explicit or accountant-based differential-privacy setting. Nine used re-identification, membership-inference or closely related disclosure testing. The remainder described privacy predominantly through architecture: raw data stayed local, faces were masked, records were deleted after processing or models ran on institutional hardware. These controls are useful, but they do not establish resistance to attacks on gradients, updates, embeddings or released synthetic data.

Security evaluation lagged behind privacy claims. Secure aggregation was explicitly integrated in only a small set of recent systems (Zhou & Yuan, 2026; Ji et al., 2026), and homomorphic or multiparty computation appeared in two primary studies (Preuveneers et al., 2021; Zhang et al., 2025). Gradient inversion, malicious-server behaviour, poisoning, collusion, sybil attacks and denial of service were seldom tested. Most FL-only studies therefore demonstrated decentralised learning, not end-to-end security.

3.5 Educational and analytical utility

Every included study reported a technical or analytical utility outcome, usually accuracy, F1, AUC, RMSE, ranking quality, fidelity or computational performance. Twenty-seven included a centralised, local or non-private comparator. The general pattern was a modest utility penalty—or, in adaptively tuned systems, occasional gains—when privacy controls were introduced. DP-TabNet, for example, achieved 80% accuracy compared with 84% for non-private TabNet (Zhao et al., 2025). Jodayree et al. (2026) reported 98.0% accuracy for FL plus DP and verifiable logging compared with 99.14% centrally. Iloh et al. (2026) found TabDDPM-generated data within 0.5 percentage points of the real-data AUC on OULAD.

High predictive performance did not always imply credible generalisation. Several studies used random or simulated client partitions, and some reported very high accuracies on curated or low-dimensional tasks (Mistry et al., 2023; Chen et al., 2026). Villegas-Ch et al. (2026) found that discrimination remained relatively stable while calibration deteriorated under institutional shift. Flanagan et al. (2022) showed

that synthetic-data performance could fail to transfer to private records. These findings make external validation, calibration and train-synthetic-test-real evaluation essential.

Educational utility was much less developed than analytical utility. No included study reported a mature, controlled multi-university deployment demonstrating improved retention, learning, feedback quality or equity attributable to the privacy-preserving system. Rejeleene et al. (2026) included faculty ratings of locally generated feedback, and several studies interpreted features for educators, but most stopped at model metrics. Consequently, the evidence supports technical feasibility more strongly than educational effectiveness.

3.6 Fairness, interpretability and reproducibility

Fairness or subgroup performance was explicitly examined in a minority of studies. Personalised FL improved predictions for demographic subgroups (Chu et al., 2022); synthetic-data work exposed privacy-fairness-utility trade-offs (Liu et al., 2025b; Kesgin, 2025); and recent federated systems added explanation stability, causal regularisation or calibration analysis (Villegas-Ch et al., 2026; Jodayree et al., 2026; Ji et al., 2026). Yet small classes and minority groups are exactly where differential noise and non-IID data may do the greatest harm, so aggregate accuracy is insufficient.

Reproducibility improved in the newest studies through public datasets, code repositories, repeated seeds and explicit attack metrics. Nevertheless, public-data reuse also created benchmark monoculture. OULAD and ASSISTments are valuable resources, but repeated optimisation against a few benchmarks cannot demonstrate robustness to different curricula, grading cultures, infrastructure, languages or legal environments.

3.7 Evidence-readiness appraisal

The descriptive appraisal showed strong reporting of technical utility but weaker reporting of deployment and assurance. All 33 studies reported a utility metric and 27 used an explicit comparator. Nineteen specified a formal privacy parameter, cryptographic protection or empirical attack test; 16 examined robustness through non-IID conditions, external datasets, holdout courses, repeated seeds or sensitivity analysis. Nine reported substantive fairness or interpretability analysis, and 14 provided code, data or unusually detailed reproducibility material. No study demonstrated a mature production federation operating across multiple autonomous higher-education institutions.

Table 4
Evidence-readiness profile of the included corpus

Appraisal domain	Studies meeting domain	Interpretation
Educational/analytical utility reported	33/33	Technical performance was universal.
Explicit comparator	27/33	Most studies permitted directional utility comparison.
Formal or attack-tested privacy evidence	19/33	Assurance was uneven; architectural claims remained common.
Robustness or external validation	16/33	Non-IID and transfer testing are increasing but not routine.
Fairness or interpretability analysis	9/33	Equity and explanation remain secondary outcomes.
Code/data/reproducibility resource	14/33	Open benchmarks help replication but narrow contextual diversity.
Live multi-institution production deployment	0/33	Institutional readiness is unproven.

4. Discussion

4.1 Principal findings

Privacy-preserving AI in digital learning has moved beyond purely conceptual proposals. Across 33 empirical studies, decentralised, noisy, synthetic, encrypted and local approaches frequently retained useful model performance. The strongest recent systems combined mechanisms rather than relying on one control—for example, federated training with differential privacy and secure aggregation, or private synthetic generation with fairness and attack evaluation. This is a meaningful advance from the shortage of applied evidence identified by Liu and Khalil (2023).

However, the field remains technically rather than institutionally mature. Most studies demonstrated feasibility on offline datasets, simulated clients or controlled partitions. Privacy was often asserted because data stayed local, even though FL can leak information through shared updates (Kairouz et al., 2021). Technical utility was commonly labelled educational utility without showing that protected analytics improved learning or decision-making. Security, fairness, governance, cost and consent were usually evaluated separately, if at all.

4.2 Decentralisation is not a privacy guarantee

The review's most important conceptual finding is the need to separate data locality from privacy assurance. FL changes where computation occurs; it does not guarantee that model updates reveal nothing about a learner. A defensible privacy claim requires a defined adversary, a protected unit (record,

learner, class or institution), measurable leakage criteria and an account of cumulative exposure across training rounds. Secure aggregation can hide individual updates from the server, while differential privacy limits inference about participation, but both require careful parameterisation and operational controls (Bonawitz et al., 2017; Dwork & Roth, 2014).

For higher education, the protected unit matters. Record-level DP may be inadequate when one student contributes hundreds of clicks, drafts and assessment events. Client-level DP may protect an institution or device but can impose more noise. Repeated model releases, dashboards and interventions also consume privacy budget or create additional inference channels. Future reports should therefore state adjacency, clipping, accounting method, epsilon, delta, number of rounds and release policy.

4.3 Benchmark success is not institutional validity

Public datasets make comparison possible, but repeated use of OULAD can conceal contextual fragility. Course structures, engagement signals, attrition mechanisms and digital infrastructure differ markedly across universities. Non-IID simulations are useful stress tests, yet they remain approximations to institutional heterogeneity, including missingness, policy differences, language, student mobility and changing curricula. External validation should include geographically and organisationally distinct institutions, temporal drift and prospective calibration.

This gap is particularly consequential for low- and middle-income settings. Bandwidth limitations, intermittent connectivity, limited GPU access, smaller class datasets and uneven data governance can alter both utility and privacy. Communication-efficient FL and CPU-capable local models are promising (Liao, 2026; Saur & Toth, 2026), but no included study established production feasibility in an African university network. A high-priority next step is a consortium-based evaluation involving public and private universities, explicit data-governance agreements and realistic infrastructure constraints.

4.4 From model utility to educational value

Accuracy is an intermediate outcome. An at-risk model has educational value only if it identifies students early enough, supports a beneficial intervention, avoids disproportionate false positives and earns legitimate trust. A private recommender has value only if its suggestions improve learning without narrowing opportunities. Future trials should therefore measure downstream outcomes such as retention, assessment performance, feedback uptake, time to intervention, educator workload, student trust and disparate impact. Privacy-preserving systems should be compared not only with centralised AI, but also with simpler data-minimising baselines and non-AI support processes.

4.5 Governance and human agency

Technical controls cannot substitute for governance. Universities need purpose limitation, data inventories, retention schedules, access control, incident response, model documentation, appeal mechanisms and human oversight. Students should know what data contribute to a model, what decisions it informs and how to challenge harmful outputs. These requirements follow established

learning-analytics ethics and are reinforced by contemporary data-protection and AI regulation (Pardo & Siemens, 2014; European Parliament & Council of the European Union, 2016, 2024; UNESCO, 2023).

5. Implications for design and institutional practice

The synthesis supports a layered minimum architecture for responsible privacy-preserving AI in higher education:

1. Minimise first: collect only variables necessary for a defined educational purpose, with retention and access rules.
2. Specify the threat model: identify trusted and untrusted parties, plausible attackers, protected units and release channels.
3. Keep raw data local where collaboration is required, but treat federated learning as an architecture rather than a guarantee.
4. Add measurable protection: use differential privacy, secure aggregation, encryption or secure computation according to the threat model.
5. Evaluate multidimensionally: report discrimination, calibration, subgroup performance, privacy leakage, robustness, communication, energy and cost.
6. Validate educationally: test whether the resulting intervention improves learning or support outcomes in prospective settings.
7. Govern continuously: document consent or lawful basis, transparency, human review, appeals, auditing and incident response.

6. Research agenda

Table 5
Priority research agenda for privacy-preserving AI in higher education

Priority	Minimum study design	Decision enabled
Real institutional federation	Prospective collaboration across at least three autonomous institutions with heterogeneous data and policies.	Whether simulated FL translates to operational settings.
Explicit adversarial evaluation	Membership and gradient-inversion tests; malicious-server, poisoning and collusion scenarios.	Whether the stated privacy/security claim survives plausible attacks.
Comparable privacy accounting	Declare adjacency, clipping, epsilon, delta, accountant, rounds and composition.	Whether privacy guarantees can be compared and audited.
Educational effectiveness	Cluster trial, stepped-wedge design or interrupted time series linking analytics to a defined intervention.	Whether protected AI improves learning, retention or support.
Equity under privacy	Intersectional subgroup metrics, small-group stability and fairness-aware calibration.	Who bears the utility cost of privacy protection.
Resource-constrained deployment	Measure bandwidth, latency, hardware, energy, staffing and total cost in LMIC institutions.	Whether the design is feasible beyond well-resourced settings.
Governance and trust	Student/educator participation, transparency testing, opt-out/appeal pathways and institutional audit.	Whether implementation is legitimate and sustainable.
Benchmark diversity	Cross-country, multilingual, longitudinal and multimodal datasets with temporal external validation.	Whether findings generalise beyond OULAD-like settings.

7. Strengths and limitations

This review integrates technical privacy, security, educational utility, fairness and deployment rather than examining any one dimension alone. It includes journal and full conference evidence from computer science, learning analytics and educational technology; distinguishes formal, empirical and architectural privacy claims; and verifies included records through DOI metadata and available full reports. The study tables make the evidence and limitations auditable.

Several limitations qualify the conclusions. First, OpenAlex and Crossref are broad discovery indexes rather than substitutes for institutional access to Scopus or Web of Science. Retrieval was capped per query family and English-language publication was required, so relevant records may have been missed. Second, screening was not independently duplicated, which increases selection risk. Third, some 2026 articles were available as accepted or advance-online versions and may change during copy-editing. Fourth, the evidence-readiness appraisal was tailored to computational educational studies and was descriptive rather than a validated risk-of-bias scale. Fifth, rapid publication means the review will require

updating. Finally, heterogeneous tasks and privacy definitions prevented meta-analysis; directional similarity in accuracy should not be interpreted as a pooled effect.

8. Conclusion

Privacy-preserving AI can support useful digital-learning analytics without routinely centralising sensitive learner records. Federated learning, differential privacy, synthetic data, secure computation and local models each showed technical promise, and combined architectures increasingly retained strong predictive utility. Yet the central evidence gap is no longer algorithmic feasibility. It is assurance and educational validity. Universities still lack evidence from live multi-institution deployments showing resistance to realistic attacks, acceptable cost, equitable performance, trusted governance and improved learner outcomes. The next generation of studies should evaluate privacy, security, utility and educational impact as a single system property. Until then, institutions should regard privacy-preserving AI as promising infrastructure requiring careful validation—not as a turnkey guarantee of responsible educational analytics.

Declarations

Funding: No external funding was received for this study.

Conflict of interest: The authors declare no conflict of interest.

Ethics approval: Not applicable. The study synthesised publicly available research and did not involve human participants or identifiable private data.

Data availability: All data supporting the synthesis are contained in the article tables and cited publications. The structured extraction matrix and search log are available from the corresponding author on reasonable request.

Author contribution: ANO: conceptualisation, methodology, investigation, evidence synthesis, writing—original draft, and writing—review and editing. MSS: evidence synthesis, writing—original draft

Declaration of generative AI use: Generative AI tools were used to support search-result organisation, metadata checking, drafting and language refinement. All cited sources were checked against DOI or publisher metadata, and the author remains responsible for the accuracy, interpretation and final submitted manuscript.

References

1. Agal, S. (2026). A privacy-preserving synthetic learner dataset for learning analytics in technology-enhanced higher education. *Scientific Reports Advance online publication*.
<https://doi.org/10.1038/s41598-026-44990-8>

2. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175–1191). <https://doi.org/10.1145/3133956.3133982>
3. Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *Bmj*, 368, l6890. <https://doi.org/10.1136/bmj.l6890>
4. Chen, S., & Qi, X. (2025). Entropy-adaptive differential privacy federated learning for student performance prediction and privacy protection: A case study in Python programming. *Frontiers in Artificial Intelligence*, 8, 1653437. <https://doi.org/10.3389/frai.2025.1653437>
5. Chen, S., Qi, X., Han, X., Fan, Z., & Wang, L. (2026). Balancing privacy and accuracy: Federated learning for multi-class student performance prediction with adaptive noise differential privacy. *Scientific Reports Advance online publication*. <https://doi.org/10.1038/s41598-026-64780-6>
6. Chu, Y. W., Hosseinalipour, S., Tenorio, E., Cruz, L., Douglas, K., Lan, A., & Brinton, C. G. (2022). Mitigating biases in student performance prediction via attention-based personalized federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 3033–3042). <https://doi.org/10.1145/3511808.3557108>
7. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407. <https://doi.org/10.1561/04000000042>
8. European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union*, L119, 1–88.
9. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
10. Fachola, C., Tornarí, A., Bermolen, P., Capdehourat, G., Etcheverry, L., & Fariello, M. I. (2023). Federated learning for data analytics in education. *Data*, 8(2), 43. <https://doi.org/10.3390/data8020043>
11. Feng, P., Zhu, H., Liu, Y., Chen, Y., & Zheng, Q. (2018). Differential privacy protection recommendation algorithm based on student learning behavior. In *2018 IEEE 15th International Conference on e-Business Engineering* (pp. 285–288). <https://doi.org/10.1109/ICEBE.2018.00054>
12. Flanagan, B., Majumdar, R., & Ogata, H. (2022). Fine grain synthetic educational data: Challenges and limitations of collaborative learning analytics. *Ieee Access : Practical Innovations, Open Solutions*, 10, 26230–26241. <https://doi.org/10.1109/ACCESS.2022.3156073>
13. Hridi, A. P., Hoq, M., Gao, Z., Lynch, C., Sahay, R., Hosseinalipour, S., & Akram, B. (2025). Privacy-preserving distributed link predictions among peers in online classrooms using federated learning. In *Proceedings of the 18th International Conference on Educational Data Mining* (pp. 276–288). <https://doi.org/10.5281/zenodo.15870179>

14. Ifenthaler, D., & Schumacher, C. (2016). Student perceptions of privacy principles for learning analytics. *Educational Technology Research and Development*, 64, 923–938.
<https://doi.org/10.1007/s11423-016-9477-y>
15. Iloh, D., Oku, G., Jiang, S., Choudhary, R., Kapa, R., & Chilakala, N. (2026). A comprehensive evaluation of generative models for privacy-preserving synthetic student data. *Ieee Access : Practical Innovations, Open Solutions*, 14, 98101–98116. <https://doi.org/10.1109/ACCESS.2026.3707491>
16. Iloh, D., Olayinka, O. T., Iyere, F., & Chilakala, S. (2025). Generative private synthetic student data for learning analytics: An empirical study. *Ieee Access : Practical Innovations, Open Solutions*, 13, 177345–177351. <https://doi.org/10.1109/ACCESS.2025.3619091>
17. Ji, W., Ying, Z., & Gan, H. (2026). BAF-FedLLM: Behavior-aware federated modeling of student actions via privacy-preserving large language model. *Mathematics*, 14(4), 604.
<https://doi.org/10.3390/math14040604>
18. Jodayree, M., Ghafi, A. K., Atashafrouz, M., & Shafiabadi, M. H. (2026). Privacy-preserving federated learning for interpretable student at-risk prediction across schools. *Scientific Reports Advance online publication*. <https://doi.org/10.1038/s41598-026-56609-z>
19. Kabir, A., & Lowd, D. (2025). On the practicality of differential privacy for knowledge tracing. In *Proceedings of the 18th International Conference on Educational Data Mining* (pp. 619–624).
<https://doi.org/10.5281/zenodo.15870248>
20. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, E., Evans, S., Gardner, D., Garrett, J., Gascón, Z., Ghazi, A., Gibbons, B., & Zhao, P. B., S (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
<https://doi.org/10.1561/22000000083>
21. Kesgin, K. (2025). FairSYN-Edu: A diffusion-based model for fair and private educational data synthesis. *Discover Education*, 4, 336. <https://doi.org/10.1007/s44217-025-00743-9>
22. Kovor, D. K., Osei, E. O., & Agyemang, E. N. (2026). Federated learning for privacy-preserving at-risk student prediction in health professional education: A multi-scenario evaluation of FedAvg and FedProx. *Computers and Education: Artificial Intelligence*, 11, 100660.
<https://doi.org/10.1016/j.caeai.2026.100660>
23. Kuzilek, J., Hlostá, M., & Zdrahal, Z. (2017). Open University Learning Analytics dataset. *Scientific Data*, 4, 170171. <https://doi.org/10.1038/sdata.2017.171>.
24. Li, X., Yan, L., Zhao, L., Martínez-Maldonado, R., & Gašević, D. (2023). CVPE: A computer vision approach for scalable and privacy-preserving socio-spatial, multimodal learning analytics. In *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (pp. 175–185). <https://doi.org/10.1145/3576050.3576145>
25. Liao, M. (2026). A privacy-preserving and communication-efficient federated learning framework for online education. *Security and Privacy*, 9(4), e70236. <https://doi.org/10.1002/spy2.70236>

26. Liu, Q., & Khalil, M. (2023). Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology*, 54(5), 1715–1747. <https://doi.org/10.1111/bjet.13388>
27. Liu, Q., Deho, O. B., Vadiiee, F., Khalil, M., Joksimović, S., & Siemens, G. (2025b). Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 591–600). <https://doi.org/10.1145/3706468.3706546>
28. Liu, Q., Khalil, M., Jovanović, J., & Shakya, R. (2024). Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 620–631). <https://doi.org/10.1145/3636555.3636921>
29. Liu, Q., Shakya, R., Khalil, M., & Jovanović, J. (2025a). Advancing privacy in learning analytics using differential privacy. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 181–191). <https://doi.org/10.1145/3706468.3706493>
30. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of Machine Learning Research*, 54, 1273–1282.
31. Mistry, D., Mridha, M. F., Safran, M., Alfarhood, S., Saha, A. K., & Che, D. (2023). Privacy-preserving on-screen activity tracking and classification in e-learning using federated learning. *Ieee Access : Practical Innovations, Open Solutions*, 11, 79315–79329. <https://doi.org/10.1109/ACCESS.2023.3299331>
32. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Bmj*, 372, n71. <https://doi.org/10.1136/bmj.n71>
33. Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450. <https://doi.org/10.1111/bjet.12152>
34. Preuveneers, D., Garofalo, G., & Joosen, W. (2021). Cloud and edge based data analytics for privacy-preserving multi-modal engagement monitoring in the classroom. *Information Systems Frontiers*, 23(1), 151–164. <https://doi.org/10.1007/s10796-020-09993-4>
35. Qin, Y., Li, M., & Zhu, J. (2023). Privacy-preserving federated learning framework in multimedia courses recommendation. *Wireless Networks*, 29(4), 1535–1544. <https://doi.org/10.1007/s11276-021-02854-1>
36. Rejeleene, R., Byram, J. N., Frankel, R. M., Isaacson, J. H., & Mehta, N. (2026). *Feasibility of privacy-preserving locally hosted language models for assessing medical student reflective writing*. *Medical Science Educator*. <https://doi.org/10.1007/s40670-026-02865-7>. Advance online publication.

37. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA Statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10, 39. <https://doi.org/10.1186/s13643-020-01542-z>
38. Rodríguez-García, M., Balderas, A., & Doderó, J. M. (2021). Privacy preservation and analytical utility of e-learning data mashups in the Web of Data. *Applied Sciences*, 11(18), 8506. <https://doi.org/10.3390/app11188506>.
39. Saur, L., & Toth, A. (2026). Grading at scale: Privacy-preserving automated short answer grading with small language models on CPUs. *Frontiers in Computer Science*, 8, 1829089. <https://doi.org/10.3389/fcomp.2026.1829089>
40. Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. <https://doi.org/10.1177/0002764213479366>
41. UNESCO (2023). Guidance for generative AI in education and research. UNESCO. <https://doi.org/10.54675/EWZM9535>
42. van Haastrecht, M., Brinkhuis, M., & Spruit, M. (2024). Federated learning analytics: Investigating the privacy-performance trade-off in machine learning for educational analytics. In *Artificial Intelligence in Education* (pp. 62–74). Springer. https://doi.org/10.1007/978-3-031-64299-9_5
43. Villegas-Ch, W., Maldonado Navarro, A., Govea, J., García-Ortiz, J., & Buenaño-Fernández, D. (2026). Federated and explainable learning analytics for privacy-preserving academic risk modeling across heterogeneous educational institutions. *Computers and Education: Artificial Intelligence*, 11, 100629. <https://doi.org/10.1016/j.caeai.2026.100629>
44. Yoneda, S., Švábenský, V., Li, G., Deguchi, D., & Shimada, A. (2025). Ranking-based at-risk student prediction using federated learning and differential features. In *Proceedings of the 18th International Conference on Educational Data Mining* (pp. 289–302). <https://doi.org/10.5281/zenodo.15870193>
45. Zhan, C., Joksimović, S., Ladjal, D., Rakotoarivelo, T., Marshall, R., & Pardo, A. (2024). Preserving both privacy and utility in learning analytics. *IEEE Transactions on Learning Technologies*, 17, 1615–1627. <https://doi.org/10.1109/TLT.2024.3393766>
46. Zhang, J., Maidin, S. S., & Dewi, D. A. (2025). BHE+ALBERT-Mixplus: A distributed symmetric approximate homomorphic encryption model for secure short-text sentiment classification in teaching evaluations. *Symmetry*, 17(6), 903. <https://doi.org/10.3390/sym17060903>
47. Zhao, Y., Wang, J., Tan, X., Wen, L., Gao, Q., & Wang, W. (2025). Privacy-preserving and interpretable grade prediction: A differential privacy integrated TabNet framework. *Electronics*, 14(12), 2328. <https://doi.org/10.3390/electronics14122328>
48. Zhou, X., & Yuan, C. (2026). AdaDP-FedSec: Adaptive differentially private federated learning with secure aggregation for multi-institutional English learner corpus collaborative training. *Scientific Reports Advance online publication*. <https://doi.org/10.1038/s41598-026-63985-z>

Figures

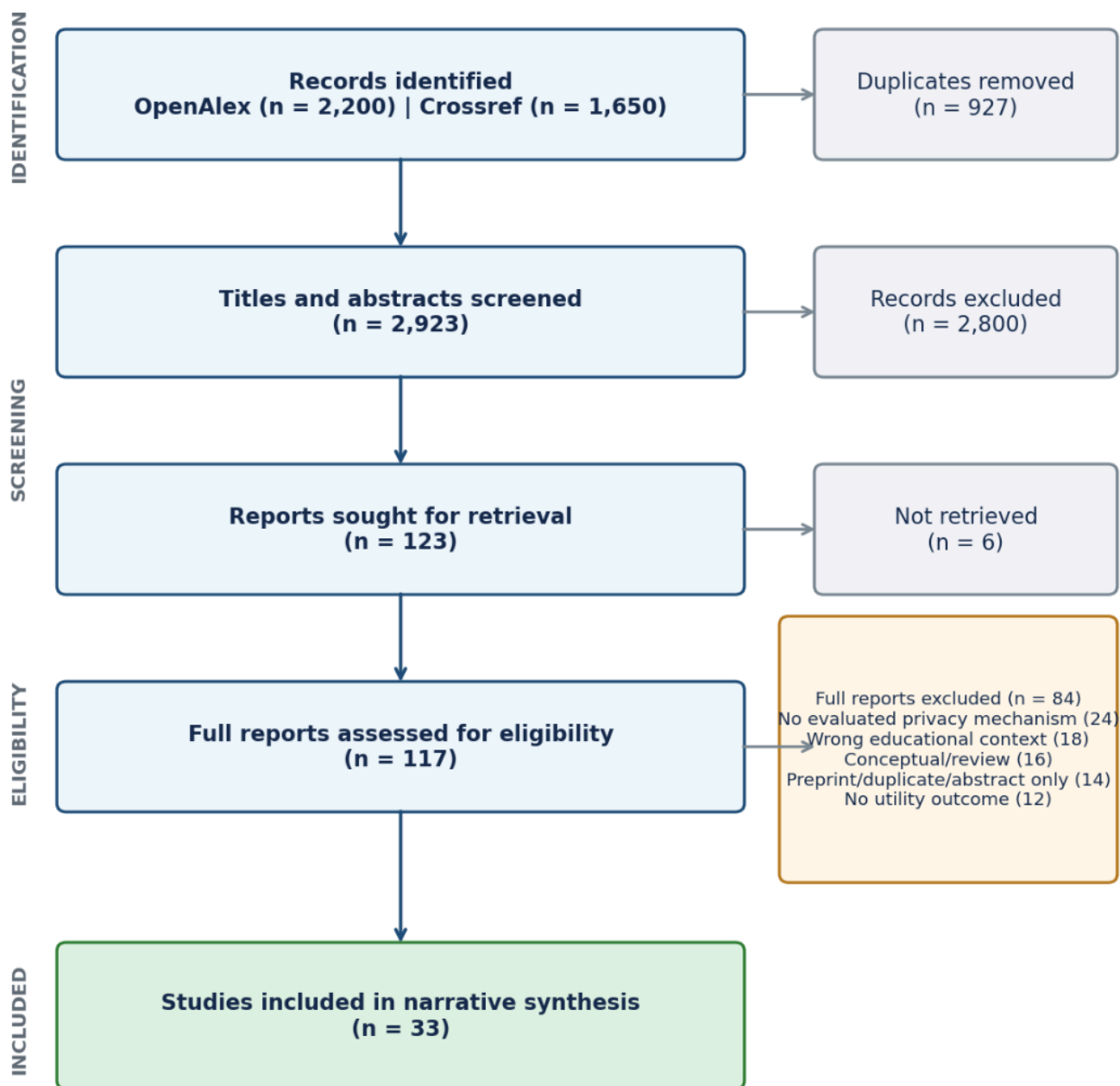


Figure 1

PRISMA 2020 flow diagram

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AppendixA.docx](#)