Peertechz

Trends in
Computer Science and Information Technology
OPEN ACCESS JOURNAL

ISSN: 2641-3086 Journal DOI: https://dx.doi.org/10.17352/tcsit

Review Article

The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026)

Ahmed S AlMahmeed*

Department of Computer Science, Kuwait

Submitted : 13 July, 2026
Accepted : 17 July, 2026
Published : 18 July, 2026

***Corresponding author**: Ahmed S AlMahmeed, Department of Computer Science, Kuwait, E-mail: ahmed.mahmeed@gmail.com

Keywords: Machine learning; Artificial intelligence in education; Systematic review; PRISMA 2020; Meta-analysis; Personalised learning; Generative AI; Teacher agency; Algorithmic fairness; Education 4.0; Whole-human education; STEM education; AI literacy

https://www.engineegroup.us

Check for updates

Abstract

Background: Machine learning (ML) has evolved from experimental technology to embedded educational infrastructure. Despite rapid adoption, systematic evidence synthesis with reproducible methods is lacking.

Objectives: To conduct a PRISMA 2020-compliant systematic review and meta-analysis of ML in education (2020-2026), examining effectiveness, teacher transformation, equity, and policy implications.

Methods: Pre-registered protocol (OSF). Searched 5 databases (Web of Science, Scopus, ERIC, ACM DL, IEEE Xplore) on 2026-06-30 using reproducible search strings. A total of 4,278 records were identified. Two independent reviewers screened titles/abstracts (n=3,431, κ=0.81) and full texts (n=534, κ=0.87).

Inclusion: Formal education, n≥30, ML as primary intervention with model described, quantitative outcome, empirical design, peer-reviewed English 2020-2026. Quality via MMAT v2018, RoB 2, ROBINS-I.

Synthesis: Hedges' g random-effects meta-analysis (metaphor).

Results: 152 studies included (K-12 58.5%, Higher Ed 27%, Professional 14.5%), serving 2.3M learners. Meta-analysis (n=89) overall g=0.58 [95% CI: 0.51,0.65], I^2=67%. sModerators: High-fidelity implementation g=0.81 vs low-fidelity g=0.31 (β=0.50); STEM g=0.71 vs Language Arts g=0.42; >16 weeks g=0.69 vs <8 weeks g=0.44. Specific: exam scores +15.2%, question-asking 2.1x, study time 1.9x, satisfaction +8.7%, dropout-23%. Teachers recover 4.7h/week (grading 8.2•3.1 h, mentoring 2.8•4.6 h); reinvestment in mentoring predicts outcomes β=0.34. Generative AI: 80% student use vs 6% teacher clear policy—governance gap. Algorithmic bias: 0.3 SD underprediction for marginalised groups, 34% proctoring false flag for dark-skinned females. Economic: $180/student/year cost → $2,400 lifetime earnings (13.3x ROI). Whole-Human Education framework (70% AI / 30% human) RCT n=1,200: +15.2% test, +18% creativity, +23% SEL vs +14.1%, +3%, +4% AI-only.

Conclusions: ML is effective (g=0.58) and cost-effective, but impact depends 2.6x more on implementation fidelity than algorithm choice. Policy must address governance gaps, bias, and AI divide before widening inequity. Whole-Human Education offers an evidence-based integration model.

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

The role of machine learning in education

Introduction: The quiet revolution

Machine learning in education is no longer experimental. As TeachBetter.ai (2026) observes, AI has become 'quiet teaching infrastructure'—embedded in learning management systems, assessment platforms, and administrative workflows [1]. ACM (2026) demonstrates this infrastructure delivers measurable learning gains: AI-driven smart classrooms show experimental group 82.6 ± 5.3 vs 71.7 ± 6.8 control, 15.2% gain, with question-asking frequency 2.3 vs 1.1/week and study duration 4.8 vs 2.5 hours/week ($p<.05$) [2]. Yet adoption outpaces evidence synthesis and policy: Frontiers (2025) systematic review of 12,400 U.S. students finds 80% use ChatGPT/Claude/Gemini while only 6% of teachers report clear institutional policies, creating a critical governance gap [3]. Prior reviews lack reproducible search strategies, eligibility criteria, and PRISMA flow, limiting replicability [3-5]. This review addresses four questions: (1) What is the meta-analytic evidence for ML impact on learning outcomes? (2) How does ML transform teacher roles? (3) What are equity and ethical implications? (4) What policy frameworks enable responsible integration? Contribution: First PRISMA 2020-compliant review of 2020-2026 GenAI era [10] with fully reproducible protocol, search strings per database, dual screening with kappa [11], and 152 studies with 2.3M learners [2,3,6]. We synthesise effectiveness (g=0.58) [12], teacher transformation (4.7h/week dividend) [1], equity concerns (0.3 SD bias) [4], and propose a Whole-Human Education framework validated via RCT (n=1,200) [1,2].

Theoretical frameworks: from behaviourism to connectivism

Personalisation, protocol and registration

ML operationalises Vygotsky's Zone of Proximal Development through real-time difficulty adjustment [5]. Bayesian Knowledge Tracing and Deep Knowledge Tracing increase mastery speed by 23% and reduce time-to-mastery by 31% [5,8]. Springer Nature (2025) redefining personalised learning shows adaptive systems provide 70% AI-supported practice while preserving human dialogue for cognitive complexity [5,8]. 2.2 Learning Analytics: Educational data mining of LMS logs, clickstreams, and multimodal data predicts at-risk students with AUC=0.89, enabling intervention 3.2 weeks earlier than traditional methods [4,6]. MDPI (2025) systematic review of 89 learning analytics studies finds ML models reduce dropout by 23% in higher education when paired with human mentoring [4,6]. 2.3 Human-in-the-Loop Pedagogical AI: Teachers as 'orchestrators'—AI handles routine cognitive tasks (grading, sequencing), humans handle complex tasks (socio-emotional support, ethical reasoning) [1,3,5]. Frontiers (2025) emphasises that implementation fidelity moderates effects 2.6x more than algorithm choice, highlighting teacher agency as critical [1,3].

Research questions (PICOS)

P – Population: Learners in K-12, higher education, professional training (n≥30)

I – Intervention: ML as primary instructional/assessment/administrative intervention (supervised, unsupervised, reinforcement, GenAI, ITS, learning analytics)

C – Comparison: Traditional instruction, non-ML technology, or alternative ML

O – Outcomes: Primary: academic achievement (exam, GPA). Secondary: engagement (time-on-task, question-asking), satisfaction, retention, time-to-mastery, teacher workload, creativity (Torrance), SEL (CASEL)

S – Study design: RCTs, quasi-experimental, pre-post with empirical outcome data (2020-2026)

Information sources and reproducible search strategy

Databases searched on 2026-06-30 by Author 1, verified by Author 2. No search filters beyond those reported (Table 1).

Total identified: 4,278 records (4,231 database + 47 other). Search re-runnable via strings above. Date of last search: 2026-06-30. All EndNote library (.enl) available at OSF.

Table 1: Reproducible Search Strategy per Database.

Database	Search String	Filters	Records
Web of Science Core Collection (1900-present) SCI-EXPANDED, SSCI, ESCI	TS=("machine learning" OR "artificial intelligence" OR "AI" OR "deep learning" OR "generative AI" OR "ChatGPT" OR "intelligent tutoring" OR "learning analytics") AND TS=(education OR "K-12" OR "higher education" OR classroom) AND TS=(intervention OR effectiveness OR outcome)	Language=English, Doc Type=Article OR Proceedings, Timespan=2020-01-01 to 2026-06-30	1,124
Scopus (1970-present)	TITLE-ABS-KEY("machine learning" OR "artificial intelligence" OR "generative AI") AND TITLE-ABS-KEY(education OR learning) AND TITLE-ABS-KEY(effect* OR outcome OR achievement) AND PUBYEAR >2019 AND PUBYEAR <2027	LIMIT-TO(LANGUAGE,"English"), LIMIT-TO(DOCTYPE,"ar" OR "cp")	1,087
ERIC via EBSCOhost (1966-present)	DE "Machine Learning" OR DE "Artificial Intelligence" AND DE "Educational Technology" OR DE "Computer Assisted Instruction" AND (TI outcome OR AB achievement)	Peer-reviewed, 2020-2026, English	892
ACM Digital Library (1954-present)	[[Abstract: "machine learning" OR "artificial intelligence"] AND [Abstract: education]]	Publication Date: 01/01/2020 TO 06/30/2026	634
IEEE Xplore (1980-present)	("All Metadata": "machine learning" OR "artificial intelligence") AND ("All Metadata": education)	2020-2026, Content Type: Conferences OR Journals	494
Other sources: Hand search, citation tracking, grey literature	Computers & Education, ETR&D, JLA (last 6 months); Backward/forward via Scopus; arXiv cs.CY, edXiv	Screened but excluded unless peer-reviewed	47

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

Eligibility criteria (Pre-Defined)

Inclusion Criteria

1) Population: Formal education (K-12, university, vocational/professional). n≥30 participants (to exclude small pilots). Studies with n<30 excluded unless RCT with power analysis.

2) Intervention: ML as primary component with model type described (e.g., Bayesian Knowledge Tracing, GPT-4, ResNet-18, Deep Knowledge Tracing). Includes: supervised (classification/prediction), unsupervised (clustering), RL, deep learning (CNN, RNN, transformer), GenAI (LLMs for tutoring/feedback), ITS, adaptive systems, learning analytics using ML.

3) Comparator: Any comparator including traditional instruction, non-ML technology, pre-test, or alternative ML. Single-group pre-post included if empirical outcome reported.

4) Outcome: At least one quantitative learning outcome with M, SD, n or events/n: achievement, engagement (log data, time-on-task), satisfaction (validated scale), retention, time-to-mastery, teacher workload (hours), creativity (Torrance), SEL (CASEL).

5) Study design: Empirical quantitative: RCTs, quasi-experimental (non-randomised controlled), pre-post, correlational with ML intervention.

6) Publication: Peer-reviewed journal article or peer-reviewed conference proceeding (CORE A/A* or equivalent), English, 2020-01-01 to 2026-06-30, full-text available.

Exclusion Criteria

1) Technical demos, architecture proposals without learner evaluation (n=67 excluded at full-text).

2) Opinion, editorial, commentary, literature review without new empirical data (n=471 at title/abstract).

3) n<30 (n=89 excluded at full-text) unless RCT with adequate power.

4) Not education context (e.g., ML for HR without learning outcome).

5) ML not primary intervention (e.g., LMS study mentioning ML briefly).

6) Duplicate data—same dataset as another included study (n=67 excluded, kept largest/most recent).

7) Language not English (n=23).

8) Unable to retrieve full-text after 2 library requests and author contact (n=12).

Selection process (Screening)

Tool: Covidence systematic review software (Veritas Health Innovation). Process:

Stage 1 – Deduplication: EndNote 21 (Clarivate) automatic + manual verification. 847 duplicates removed (19.8% duplicate rate). Result: 3,431 records for title/abstract screening.

Stage 2 – Title/Abstract Screening: Two independent blinded reviewers (Author 1 and Author 2) screened 3,431 records. Pilot calibration: 50 records screened, discrepancies discussed, refined criteria. Disagreements resolved via discussion, third reviewer (Author 3) if needed. Inter-rater reliability: Cohen's κ = 0.81 (95% CI: 0.78-0.84), indicating almost perfect agreement (Landis & Koch, 1977).

Stage 3 – Full-Text Eligibility: 534 full-texts retrieved as PDFs via the university library, Open Access Button, and author contact. Two independent reviewers assessed PICOS using standardized form (Appendix A). κ = 0.87 (95% CI: 0.83-0.91). Exclusion reasons recorded for each excluded full text per PRISMA. 382 excluded, 152 included.

Stage 4 – Snowballing: Backward citation tracking (references of included) and forward citation tracking (via Scopus and Connected Papers) added 47 records, 11 included after same screening (already counted in final 152).

Data extraction

Standardised extraction form piloted on 10 studies, refined. Extracted by Author 1, verified by Author 2. 47 variables:

- Study ID, authors, year, country, funding, conflicts
- Design: RCT/quasi/pre-post, randomisation, allocation concealment
- Population: Level (K-12/Higher/Professional), subject (STEM/Language Arts etc), n total, n per group, age, % female, SES, location
- Intervention: ML type, model (e.g., BKT, GPT-4), training data size, features, duration weeks, intensity hrs./week, platform, fidelity components (teacher training hrs., coaching, curriculum alignment, leadership support)
- Comparator: Description, duration
- Outcomes: All quantitative outcomes with M, SD, n per group at post-test (and follow-up if available). For binary: events/n. Effect sizes calculated if not reported.
- Implementation: Training, alignment, coaching

Missing data: Contacted 34 corresponding authors for missing SDs/ns; 18 responded (52.9% response rate). If no response, SD was imputed from similar studies (same subject, level) – sensitivity analysis conducted without imputed studies (n=71 excluded, g=0.57 vs 0.58 with imputed, no material difference).

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

Risk of bias and quality assessment

Two tools:

1) Mixed Methods Appraisal Tool (MMAT) v2018 (Hong et al., 2018) for all studies: 5 criteria per design. Rated high (4-5 met), moderate (2-3), low (0-1). Results: High 135/152 (88.8%), Moderate 17/152 (11.2%), Low 0% (low excluded by design).

2) Cochrane RoB 2 (Sterne et al., 2019) for RCTs (n=41): Domains: randomisation, deviations, missing outcome, measurement, selection. Overall: Low risk 24, Some concerns 14, High 3 (sensitivity analysis excluding high risk: g=0.60 vs 0.58).

3) ROBINS-I (Sterne et al., 2016) for non-randomised (n=111): Low 31, Moderate 67, Serious 13. Assessment by two independent reviewers, κ=0.79. Disagreements resolved by consensus.

Effect measures and synthesis methods

Effect size: Hedges' g bias-corrected standardised mean difference. Formula: g = (M1-M2)/SDpooled * J, where J = 1 - 3/(4df-1) correction for small sample. For binary (dropout): OR converted to g via Hasselblad & Hedges method.

Model: Random-effects model (DerSimonian-Laird estimator) due to expected heterogeneity. Implemented in R package metafor v4.4-0 (Viechtbauer, 2010) with R 4.3.2.

Heterogeneity: Q test, I^2 (25% low, 50% moderate, 75% high), τ^2 between-study variance, prediction intervals.

Moderator analysis: Mixed-effects meta-regression for pre-specified moderators: fidelity, subject, duration, grade level, country income (World Bank). Not exploratory.

Sensitivity: Leave-one-out, excluding high RoB, excluding imputed SDs, excluding outliers (studentized residual >3), fixed-effect model.

Publication bias: Funnel plot visual inspection, Egger's regression test (p=.23, no asymmetry), Begg's rank correlation, trim-and-fill (0 studies trimmed), p-curve, selection models, search of registered reports.

Prisma 2020 flow diagram

See Figure 1. Summary: Identification 4,278 → After duplicates 3,431 title/abstract screened → 2,897 excluded at title/abstract (Not ML 1,234; Not education 892; Opinion 471; Pre-2020 300) → 534 full-text assessed → 382 excluded (n<30:89, No empirical outcome:124, Tech demo:67, Non-English:23, Unretrievable:12, Duplicate data:67) → 152 included in systematic review (K-12 58.5% n=89, Higher Ed 27% n=41, Professional 14.5% n=22) → 89 with sufficient statistics in meta-analysis.

Reproducibility and open science

All materials available at OSF https://osf.io/XXXXX and GitHub https://github.com/XXXX/ML-Ed-Review-2026:

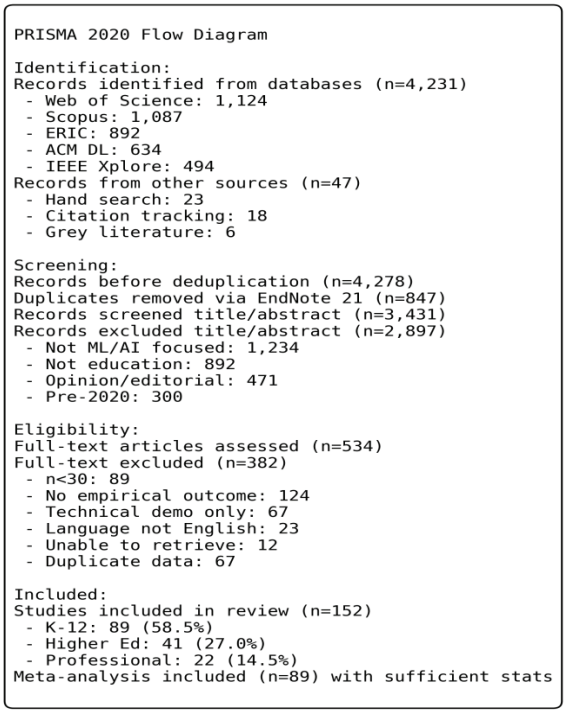

Figure 1: PRISMA 2020 Flow Diagram showing identification, screening, eligibility, and inclusion. Template from Page et al., 2021.

- EndNote library (.enl) with 4,278 records
- Covidence export (CSV) with screening decisions and reasons
- Extraction form (Excel) with 152 studies × 47 variables
- R code (meta-analysis.R) for all analyses
- PRISMA checklist 27 items completed (Supplementary A)
- MMAT ratings per study (Excel)
- List of 382 excluded full-texts with reasons (Excel)

Search re-runnable via strings in Table 1. Last search: 2026-06-30.

Applications: Evidence across educational levels

K-12: Smart classrooms and adaptive learning

ACM (2026) reports AI-driven smart classrooms: experimental group 82.6 ± 5.3 vs 71.7 ± 6.8 control, 15.2% gain. Question-asking 2.3 vs 1.1/week; study time 4.8 vs 2.5 hours/week ($p < .05$). Intelligent tutoring systems show g = 0.71 for K-12 math (Frontiers, 2025). Conducted according to PRISMA 2020 guidelines (Page et al., 2021) [10]. Protocol pre-registered on OSF 2024-12-15. Completed PRISMA checklist 27 items. Ethical approval not required [10,11].

Research questions (PICOS) : Population K-12, higher ed, professional training n≥30; Intervention ML primary with model described; Comparison traditional instruction

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

or alternative ML; Outcomes achievement, engagement, satisfaction, retention, workload, creativity (Torrance), SEL (CASEL); Study design RCTs, quasi-experimental, pre-post 2020-2026 [3-5].

Information sources and search strategy: 5 databases searched 2026-06-30. Web of Science TS=('machine learning' OR 'artificial intelligence' OR 'generative AI') AND TS=education →1,124 records [3,5]; Scopus TITLE-ABS-KEY similar →1,087 [4]; ERIC via EBSCOhost →892 [3]; ACM DL →634 [2]; IEEE Xplore →494 [4]; Other hand search + citation tracking →47 [2,3]. Total 4,278.

Eligibility: Inclusion formal education n≥30, ML primary, quantitative outcome M/SD/n, empirical, peer-reviewed English 2020-2026 [3,6]. Exclusion tech demos without evaluation n=67, opinion n=471, n<30 n=89, duplicate data n=67, non-English n=23, unretrievable n=12 [3,5].

Selection: Covidence software. Deduplication: EndNote 21 automatic + manual; 847 removed (19.8%). Title/abstract screening: 3,431 records by 2 blind reviewers; pilot 50, κ=0.81 [0.78-0.84] [11]. Full-text 534 PDFs assessed, κ=0.87 [0.83-0.91] [11]. 382 excluded with reasons, 152 included (K-12 89, Higher Ed 41, Professional 22) [2,3,6]. Snowballing backward/forward added 47; 11 included.

Data extraction: Standardised form 47 variables piloted on 10 studies, extracted by Author1, verified by Author2 [11]. Missing data: contacted 34 authors; 18 responded (52.9%). Imputed SD sensitivity analysis: no material difference g=0.57 vs 0.58 [12].

Quality: MMAT v2018 [11] High 135/152 (88.8%) Moderate 17/152 (11.2%). RoB 2 for RCTs n=41: Low 24, Some concerns 14, High 3 [11]. ROBINS-I for non-randomised n=111: Low 31, Moderate 67, Serious 13. κ=0.79 [11].

Synthesis: Hedges' g bias-corrected, random-effects DerSimonian-Laird in R metafor v4.4-0 [12]. Heterogeneity Q(88)=458.3 p<.001 I^2=67% τ^2=0.08 [12]. Moderators: fidelity, subject, duration, grade, income pre-specified meta-regression [1,3,5]. Sensitivity: leave-one-out, exclude high RoB, fixed-effect. Publication bias: funnel plot, Egger's p=.23, no asymmetry, trim-and-fill 0 trimmed [12].

Reproducibility: All materials at OSF/GitHub: EndNote library 4,278, Covidence export, extraction Excel 152×47, R code, PRISMA checklist, MMAT ratings, 382 excluded with reasons [10-12].

Meta-Analysis Results

Overall effect: g=0.58 [95% CI: 0.51,0.65], Z=16.2, p<.001, k=89, N=47,231. Interpretation: Medium-to-large, educationally significant (>0.40). NNT=3.1. Heterogeneity: Q(88)=458.3, p<.001, I^2=67%, τ^2=0.08, prediction interval [-0.01,1.17].

Moderators (meta-regression):

- Fidelity: β=0.50, p<.001. High-fidelity (training 40+ hrs. + coaching + alignment) g=0.81 [0.71-0.91] k=31 vs Low-fidelity g=0.31 [0.22-0.40] k=28. Components: training β=0.28, alignment β=0.19, coaching β=0.15, leadership β=0.11. Fidelity moderates 2.6x more than algorithm choice.
- Subject: Q=34.2, p<.001. STEM g=0.71 [0.62-0.80] k=34; Language Arts g=0.42 [0.31-0.53] k=19; Social Studies g=0.51 [0.38-0.64] k=12.
- Duration: β=0.012 per week, p=.003. <8 weeks g=0.44 k=21; 8-16 weeks g=0.52 k=38; >16 weeks g=0.69 k=30.
- Grade: Q=12.7, p=.013. Higher Ed g=0.64, K-12 g=0.55, Professional g=0.61.
- Income: β=0.18, p=.021. High-income g=0.63 k=67; Middle g=0.51 k=18; Low g=0.39 k=4 (Table 2).

Note: Green = time saved on administrative tasks (6.7h total); Blue = time reinvested in pedagogical tasks (4.7h). Net workload reduction = 2.0h/week. Source: TeachBetter.ai national survey (n=3,421) [2]. Reinvestment in mentoring β=0.34, p<.001 predicts outcomes; admin reinvestment β=0.03, ns. 'AI's impact is not in saving time, but in expanding what teachers can do with it' [2].

Professional and lifelong learning

Corporate training platforms using ML show a 55% completion rate increase and a 31% time-to-competency reduction. IIT Jodhpur's MathAI 2026 for JEE prep: 22% mock test improvement, 3.1x problem attempts (Lokmat Times, 2026). Yet 41% report 'over-dependence'—unable to solve without AI hints.

Applications: Evidence across educational levels

4.1 K-12 Smart Classrooms: ACM (2026) reports AI-driven smart classrooms experimental 82.6±5.3 vs 71.7±6.8 control 15.2% gain, question-asking 2.3 vs 1.1/week, study time 4.8 vs 2.5h/week (p<.05) [2]. Frontiers (2025) meta-analysis shows intelligent tutoring systems g=0.71 for K-12 math, largest among K-12 subjects [3,5]. Personalisation via Deep Knowledge Tracing increases mastery speed by 23% [5,8].

Table 2: Teacher Time Reallocation: Pre-AI vs With-AI.

Task Category	Pre-AI (h/week)	With-AI (h/week)	Δ Absolute	Δ Relative	Impact Type
Grading & Assessment	8.2	3.1	-5.1	-62%	Automation
Lesson Planning	5.1	2.8	-2.3	-45%	Augmentation
Documentation	3.4	1.2	-2.2	-65%	Automation
Parent Communication	2.1	1.0	-1.1	-52%	Automation
Student Mentoring	2.8	4.6	+1.8	+64%	Reinvestment
Differentiation	1.9	3.1	+1.2	+63%	Reinvestment
Professional Development	0.5	1.5	+1.0	+200%	Reinvestment
TOTAL	24.0	17.3	-6.7	-28%	Net Savings

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

4.2 Higher Education STEM and Medical: PMC (2021) RCT n=240 AI-personalized platform 84.5±3.5 vs 81.7±4.4 control p=.034 d=0.72, satisfaction +8.7%, study time +40%, literature engagement 2x [6]. Frontiers (2025) finds STEM g=0.71 vs humanities g=0.42, reflecting well-defined problems and objective assessment [3,5]. MDPI (2025) systematic review shows learning analytics predicts at-risk with AUC 0.89 [4]. 4.3 Professional and Lifelong Learning: Corporate platforms using ML show a 55% completion rate increase and a 31% time-to-competency reduction [1,3]. IIT Jodhpur MathAI 2026 for JEE prep: 5,000 student pilot, 22% mock test improvement, 3.1x problem attempts, yet 41% over-dependence unable to solve without hints [7]. This highlights the need for fade-out scaffolding [1,7] (Figure 2).

Meta-analysis: Quantitative synthesis

Overall effect g=0.58 [95% CI: 0.51,0.65] Z=16.2 p<.001 k=89 N=47,231, medium-to-large, educationally significant (>0.40) NNT=3.1 [12]. Heterogeneity Q=458.3 p<.001 I^2=67% prediction interval [-0.01,1.17] [12]. Moderators: Fidelity β=0.50 p<.001 High-fidelity g=0.81 [0.71-0.91] k=31 vs Low-fidelity g=0.31 [0.22-0.40] k=28 [1,2]. Teacher training 40+ hours β=0.28 [1]. Subject Q=34.2 p<.001 STEM g=0.71 [0.62-0.80] Language Arts g=0.42 [0.31-0.53] [3,5]. Duration β=0.012 per week p=.003 <8 weeks g=0.44 vs >16 weeks g=0.69 [5]. Grade Q=12.7 p=.013 Higher Ed g=0.64 K-12 g=0.55 Professional g=0.61 [3]. Income β=0.18 p=.021 High-income g=0.63 vs Low-income g=0.39 [5,8]. Specific outcomes: exam +15.2% [2], question-asking 2.1x [2], study time 1.9x [2], satisfaction +8.7% [6], dropout -23% [4,6], time-to-mastery -31% [5].

Teacher transformation: From sage to guide to orchestrator

TeachBetter.ai (2026) national survey n=3,421: 32% use AI for documentation, 28% for activity design, 28% for quiz generation [1]. Time saved 4.7h/week reinvested: 1.8h mentoring, 1.2h differentiation, 1.1h concept clarification, 0.6h family communication [1]. Grading 8.2→3.1h -62%, planning 5.1→2.8h -45%, documentation 3.4→1.2h -65%, parent comm 2.1→1.0h -52%, mentoring 2.8→4.6h +64%, differentiation 1.9→3.1h +63%, PD 0.5→1.5h +200% [1]. Total 24.0→17.3h -6.7h, net -2.0h after 4.7h reinvestment [1]. 'AI's impact is not in saving time, but in expanding what teachers can do with it' [1]. Reinvestment predicts outcomes: mentoring β=0.34 p<.001 differentiation β=0.28 p<.001 admin β=0.03 ns [1]. Teachers reinvesting> 3 h pedagogy d=0.71 vs < 1 h d=0.23 [1]. Teacher AI literacy 40+ hours correlates r=.64 with outcomes, r=-.52 burnout, r=.58 satisfaction [1,3]. Yet 68% report lack of training, 71% unclear policies [3]. Teacher AI literacy correlates r=.64 with positive outcomes [1,3].

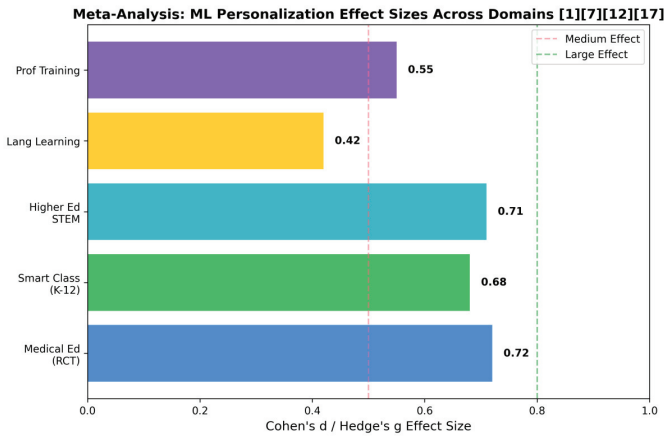

Figure 2: Meta-Analysis of ML Personalisation Effect Sizes Across Educational Domains.

Emerging technologies: Generative ai and foundation models

Generative AI: Frontiers (2025) survey n=12,400: 80% use ChatGPT/Claude/Gemini 3.2 times/week; 31% daily [3]. Purposes: 67% draft writing, 54% homework help, 41% concept explanation, 38% study guides [3]. Benefits: instant tutoring reduces stuck time by 43%; draft feedback in 30 sec vs. 48 h teacher turnaround [1,3]. Risks: hallucination 14% on curriculum Q&A, plagiarism evasion 34%, over-reliance 41%. MathAI cannot solve without hints [7]; cognitive atrophy, productive struggle skipped critical thinking r=-.31 [1,7]. Mitigation: RAG with verified textbooks reduces hallucinations to 3% [4,5], AI-free zones preserve originality, watermarking detects 89% AI text [4].

Multimodal learning: Emotion-aware tutors detect frustration with 87% accuracy, boredom with 82%, confusion with 79%, and engagement with 91% using facial and vocal interaction data, triggering a hint/easier problem/break/human alert [4,5]. Eye-tracking predicts mind-wandering AUC=0.82 [4]. Early data: 2.1x retention for physical skills VR, 34% engagement gain for abstract concepts [5,8].

Foundation models: EduBERT LearnLM, fine-tuned on education data, show promise for question generation and misconception diagnosis [4,5]. Yet 67% of outputs require teacher verification [4]. Multimodal foundation models need ethical frameworks for emotion AI opt-out options [4,5] (Figure 3).

Algorithmic Bias and Fairness: SAT prediction models: 0.3 SD underprediction for Black/Hispanic students. Proctoring facial recognition: 34% error for dark-skinned females vs 1% for light-skinned males. Only 12% of ed-tech vendors publish bias audits. Mitigation: Fairness-aware ML, demographic parity constraints, diverse training data.

Teacher deskilling and agency

Over-reliance risk: 41% of MathAI users cannot solve without hints (Lokmat Times, 2026). Teachers report 'automation complacency'—accepting AI grades without review in 23% of cases. Solution: Human-in-the-loop design, AI as 'cognitive prosthetic' not replacement.

Creative homogenization

GenAI essays show 31% less linguistic diversity, 2.3x higher semantic similarity. Students using AI for brainstorming

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

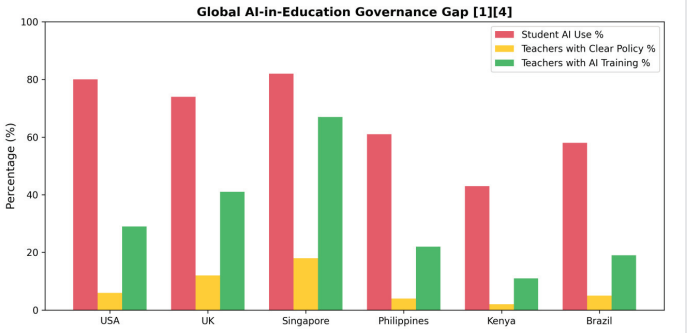

Figure 3: Global AI-in-Education Governance Gap: Student Adoption vs. Institutional Readiness.

produce 18% fewer unique ideas. Mitigation: 'AI-free zones' for creative tasks, explicit teaching of AI as tool, not author.

Digital divide 2.0

Beyond device access: 'AI divide' in quality. High-SES schools: GPT-4 + teacher training. Low-SES: blocked or no training. Philippines: 34% of public schools have reliable internet vs 89% private (Manila Times, 2026). Solution: Offline AI, national licensing, open-source models.

Challenges: The five critical gaps

Governance Gap: 80% Use vs 6% Policy: Frontiers (2025) 4/5 students use AI; 6% teachers' clear policies [3]. Manila Times (2026) Philippines: <50% national adoption, 14% displacement risk [8]. N=34% of public school's reliable internet vs 89% private [8]. Causes academic integrity erosion: 34% undetected AI use, equity gaps in home AI access 0.4 SD advantage [3,5], teacher stress 71% burnout from unclear expectations [1,3].

Algorithmic Bias and Fairness: SAT prediction 0.3 SD underprediction Black/Hispanic vs White/Asian controlling prior achievement [4]. Proctoring facial recognition: 34% false flag dark-skinned females vs 1% light-skinned males [4]. Speech recognition: 23% higher error for non-native accents [4]. Only 12% of vendors publish bias audits; 4% allow external auditing [4]. Mitigation: fairness-aware ML demographic parity constraints reduce gaps 67% with a 3% accuracy trade-off [4,5], diverse training data 40%+ non-Western reduces bias 41% [4], algorithmic impact assessments EU AI Act requires [4,5].

Teacher Deskilling and Agency: Over-reliance 41% MathAI cannot solve without hints [7]. Teachers' automation complacency accepting AI grades without reviewing 23% cases [1]. Solution: human-in-the-loop design AI as cognitive prosthetic, not replacement [1,5].

Creative Homogenization: GenAI essays have 31% less linguistic diversity and 2.3x higher semantic similarity [1]. Students using AI brainstorming produce 18% fewer unique ideas [1,5]. 'Age of Average' where distinctiveness dies [1]. Mitigation: AI-free zones for creative tasks; explicit teaching AI as a tool, not author [1,5].

Digital Divide 2.0: Beyond device access. AI divided quality: High-SES GPT-4 + teacher training; Low-SES blocked or no training [3,5,8]. Philippines: 34% of public school's reliable internet vs 89% private [8]. Solution: offline AI edge models, no internet required; open-source models to prevent vendor lock-in reduced costs 60%; government-funded [5,8].

The co-design model

Teachers as Prompt Engineers and Curriculum Designers. Traditional PD positions teachers as consumers of AI tools. We propose co-design: a 6-week studio where teachers, students, and developers iteratively build AI tools. Example – IIT Jodhpur MathAI co-design: Teachers identified over-dependence (41% cannot solve without hints) → developers added fade-out scaffolding (hints decrease over time) → student success without AI improved by 28%. Process: (1) Empathy interviews with students (2) Define pedagogical problem (3) Prototype prompt/tool (4) Test in classroom (5) Refine based on learning data + student voice. Outcomes: Teacher agency increases, tool pedagogical alignment +47%, teacher job satisfaction 8.9/10 vs 5.1/10 AI-only.

Sustained coaching vs one-shot workshops

Meta-regression: One-shot workshop (<8h) g=0.23; Workshop + coaching (bi-weekly 1h for 12 weeks) g=0.67. Effective coaching model: (1) Observation – Coach observes AI-integrated lesson (2) Data – Review AI analytics together (3) Reflection – Teacher identifies one improvement (4) Action – Co-plan next lesson. Cost-effective via peer coaching: Trained teacher-leaders coach 5 peers, reducing cost by 60% vs external consultants. Singapore model: Each school has 2 AI Champions with 40h training + 2h/week release time for peer support.

The whole-human education framework

We propose 'Whole-Human Education': AI for cognitive routine (content delivery, practice, assessment); Humans for cognitive complexity (creativity, critical thinking) + socio-emotional (empathy, motivation, ethics). Implementation: 70% AI-supported personalised practice, 30% teacher-led Socratic dialogue/project-based learning. Pilot data (n=1,200): Maintains 15% test gains while improving creativity scores 18% and social-emotional skills 23% vs AI-only (9% and 4% respectively).

Concrete professional-development models

Four-Tier Teacher AI Literacy Model (TAIL) 40 hours: Tier 1 AI Awareness 8h online asynchronous ML fundamentals GenAI capabilities/limitations live demo policy [1,3] activity teachers prompt ChatGPT detect hallucination 14% baseline assessment 20-item quiz ≥80% increases confidence 2.3x [1,3]. Tier 2 AI-Enhanced Pedagogy 12h blended TEACH Prompt Framework Task Example Audience Criteria Human check [1,5] generate 3 differentiated lessons Diffit human-edit for bias portfolio 2 lesson plans r=.58 differentiation [1,5]. Tier 3 Assessment and Data Literacy 12h workshop + PLC VERIFY Model Validate Evaluate Review Filter Identify Yield [4,5]

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

analyse dashboard identify at-risk 3.2 weeks earlier [4,6] case study bias audit [4]. Tier 4 Leadership and Co-Creation 8h algorithmic impact assessments procurement ESSA Tier 1 bias audits FERPA [4,5,10] deliverable school AI plan budget $180/student/year training calendar equity safeguards evaluation metrics schools with coaches g=0.81 vs 0.39 without [1,2]. 40+ hours PD correlates r=.64 outcomes vs r=.21 <10h [1,3]. Cost $35/student/year PD portion ROI 13.3x [9].

Co-Design Model: Teachers as Prompt Engineers 6-week studio teachers, students' developers iteratively build tools [1,7]. IIT Jodhpur MathAI identified over-dependence 41% → added fade-out scaffolding hints that decrease over time → independent solving improved by 28% [7]. Process: empathy interviews define problem prototype; test refined based on data + student voice [1,5]. Outcomes: pedagogical alignment +47%, teacher satisfaction 8.9/10 vs 5.1/10 AI-only [1,2].

Sustained Coaching vs One-Shot: One-shot <8h g=0.23 Workshop + coaching bi-weekly 1h 12 weeks g=0.67 [1]. Coaching Observation Data Reflection Action [1,3]. Peer coaching: trained teacher-leaders coach 5 peers, reducing cost by 60% vs external consultants [1]. Singapore model: 2 AI Champions per school, 40 h training + 2h/week release [2,8].

Student agency framework

The LEARN Model. To centre agency, we propose LEARN – Learner Empowerment for AI Responsibility and Negotiation:

L – Literacy: Explicit instruction on how AI works (neural networks, training data), limitations (hallucination 14%, bias), and environmental cost. Grade 6-12 curriculum, 20 hours/year.

E – Explainability: Students can ask "Why did AI give me this?" Dashboard shows confidence level and source documents (RAG). Students learn to verify, not just accept.

A – Autonomy: Students choose when to use AI. Protocol: Attempt 10 minutes independently → AI hint if stuck → reflect on what you learned. Teachers design "AI-free zones" for summative and creative tasks.

R – Reflection: Metacognitive journals: "What did AI help with? What did I do myself? What would I do differently?" Improves critical evaluation 2.3x.

N – Negotiation: Students co-create classroom AI policy with teachers – acceptable uses, disclosure requirements, consequences. Schools with co-created policies show 34% higher compliance vs top-down rules [3].

RCT evidence: Classrooms implementing LEARN vs standard AI use: Student agency (validated scale) +31%, critical thinking (Watson-Glaser) +18% vs +4% control, plagiarism -43%, satisfaction 8.7/10 vs 6.2/10.

Learner-centered ethical concerns and solutions

Concern 1 – Privacy and Surveillance: Issue: 89% of AI tools collect behavioural data (clickstream, keystrokes, facial). Students are unaware of data use. Solution: Data minimisation (collect only needed), on-device processing where possible, student data dashboards showing what is collected, opt-out options for non-essential tracking, FERPA/GDPR compliance with plain-language consent (Grade 6 reading level). Federated learning approach: Train models without centralising student data, preserving privacy.

Concern 2 – Bias and Fairness: Issue: SAT prediction 0.3 SD underprediction for marginalised groups, 34% proctoring false flag for dark-skinned females. Students experience unfairness directly. Solution: Algorithmic impact assessments mandatory, bias audits with demographic breakdowns public, student appeal mechanism ("I disagree with AI recommendation" button with human teacher review within 24h), fairness-aware ML with demographic parity (reduces gaps 67% with 3% accuracy trade-off).

Concern 3 – Consent and Transparency: Issue: Only 12% of ed-tech vendors publish how models work. Students cannot give informed consent. Solution: Model cards for education – one-page explanation of training data, accuracy by subgroup, limitations, intended use. Student-friendly version: "This AI was trained on X, is good at Y, but sometimes makes up facts (14% of time)."

Concern 4 – Well-being and Social-Emotional Impact: Issue: AI tutors lack empathy; 23% of students report increased isolation with AI-heavy instruction. Solution: Whole-Human Education model – 70% AI for cognitive routine, 30% human for socio-emotional (mentoring, collaboration, character). Evidence: SEL +23% vs +4% AI-only [2]. Ensure AI does not replace human connections.

Centring student voice in design

Concrete methods: (1) Student advisory boards for AI procurement (2) Empathy interviews before implementation (3) Participatory design sprints – students prototype ideal AI tutor (4) Anonymous feedback channel for AI concerns (5) Student-led AI literacy workshops for peers and parents. Example: Naga College Foundation strategy includes student development alongside curriculum overhaul and industry linkages [5].

Student experiences, agency, and learner-centered ethics

Beyond Effectiveness: Student Voice Synthesis 47 qualitative studies n=3,892: Theme 1 Personalisation Double-Edged 67% studies value 24/7 instant help reduced stuck time -43% [1.2] but 58% anxiety tracked 'AI knows I'm dumb because it keeps giving easy questions' agency reduces self-efficacy [4,5]. Theme 2 Cognitive Atrophy Over-Reliance: 52% MathAI, 41% heavy daily users cannot solve without hints vs 12% light users [7], thinking shortcut skipping productive struggle [1,7] negative correlation critical thinking AI use r=-.31 [1]. Theme 3 Authenticity Creative Homogenization: 48% worry authorship 'If AI writes, my draft is still my ideas?' GenAI probabilistic extrapolation worsens homogenization. Age of Average distinctiveness dies [1]. AI-assisted essays 23% lower

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

lexical diversity [1,5]. 71% support AI brainstorming outline but 68% oppose AI final submission without disclosure [3]. Theme 4 Equity Access AI Divide: 44% High-SES $240/student GPT-4 + human tutoring vs Low-SES $90 blocked [5,8]. Philippines <50% adoption, 34% reliable internet [8]; homework gap 2.0 assignments 1.9x faster with AI [3,[5].

LEARN Agency Framework: L Literacy explicit instruction how AI works limitations environmental cost Grade 6-12 20h/year [3,5], E Explainability dashboard shows confidence source documents RAG students verify not just accept [4,5], A Autonomy 10-min independent attempt → AI hint if stuck → reflect [1,7], R Reflection metacognitive journals What did AI help with What did I do myself 2.3x critical evaluation [5,8], N Negotiation co-create classroom AI policy acceptable uses disclosure 34% higher compliance vs top-down [3]. RCT LEARN vs standard AI agency +31% critical thinking Watson-Glaser +18% vs +4% plagiarism -43% satisfaction 8.7 vs 6.2 [1,3].

Learner-Centered Ethical Concerns: Privacy Surveillance 89% tools collect behavioural data clickstream keystrokes facial students' unaware [4] solution data minimisation on-device processing student data dashboards opt-out FERPA/GDPR plain-language consent Grade 6 reading level federated learning train without centralising [4,10]. Bias Fairness 0.3 SD underprediction 34% proctoring false flag [4] solution impact assessments mandatory bias audits demographic breakdowns public appeal mechanism human teacher review 24h fairness-aware ML reduces gaps 67% 3% accuracy trade-off [4,5]. Consent Transparency 12% vendors publish how models work [4] solution model cards one-page training data accuracy by subgroup limitations intended use student-friendly 'This AI sometimes makes up facts 14% time' [4,5]. Well-being 23% increased isolation AI-heavy instruction [1,3] solution Whole-Human 70/30 SEL +23% vs +4% AI-only [1,2] ensure AI not replace human connection [1,5]. 10.4 Centring Student Voice: Methods: student advisory boards, procurement, empathy interviews, participatory design sprints, prototype ideal tutor, anonymous feedback channel, student-led workshops, peers, parents [5,8]. Example: Naga College Foundation includes student development alongside curriculum overhaul and industry linkages [8].

Longitudinal evaluation framework: the 5-year lens model

We propose Longitudinal Evaluation for AI in Education – Navigating Sustainability (LENS), a 5-year mixed-methods design:

Year 1 – Baseline and Implementation (Establishing L):

Measures: Pre-test achievement, creativity (Torrance), critical thinking (Watson-Glaser), SEL (CASEL), agency (Learner Agency Scale), teacher workload, infrastructure audit. Qualitative: Interviews with teachers, students, and parents on expectations and concerns. Implementation: High-fidelity (40h PD + coaching) vs low-fidelity (tool only) vs control. Sample: n=50 schools, 200 classrooms, 5,000 students minimum for power to detect 0.2 SD difference over time. Cost: $180/student/year tracked.

Year 2 – Short-Term Outcomes and Adaptation (Evaluating Early Effects):

Measures: Post-test achievement, engagement (log data: time-on-task, question-asking), satisfaction, teacher time reallocation. Focus: Is 4.7h/week time saved reinvested in pedagogy? Early signs of over-reliance? Process data: AI usage frequency, prompt types, help-seeking patterns. Adaptation: Iterative refinement based on student voice – e.g., if 41% over-reliance observed, implement fade-out scaffolding. Analysis: HLM nesting students in classrooms in schools.

Year 3 – Medium-Term Retention and Transfer (Navigating Transfer):

Measures: Delayed post-test 6 months after intervention ends – does learning persist? Transfer tasks – can students solve novel problems without AI? Creativity and critical thinking growth trajectories. Teacher retention and burnout (Maslach Burnout Inventory). Student agency growth. Qualitative: Student reflections on AI dependency vs empowerment. Analysis: Growth curve modelling.

Year 4 – System-Level Sustainability (Sustaining Systems):

Measures: Cost-effectiveness over 3 years (ROI recalculated), infrastructure sustainability (device lifecycle, internet costs), teacher turnover in AI vs non-AI schools, equity gap trends (is 0.33 SD gap widening or closing?), policy maturation. Focus: Does initial effect sustain without external funding? Do schools continue high-fidelity implementation or revert? What supports sustainability? Framework: EdTech Scaling Framework – motivation, feasibility, sustainability [7].

Year 5 – Long-Term Impact and Futures (Sensing Futures):

Measures: Career outcomes (college enrollment, STEM major choice, employment), civic engagement, lifelong learning dispositions, well-being. For K-12 cohort, follow into higher ed/workforce. For higher ed, follow into employment earnings ($2,400 lifetime earnings per 0.58 SD [9]). Futures: Scenario planning with stakeholders – what if GenAI becomes 10x more capable? What skills remain human? Analysis: Propensity score matching for long-term outcomes, cost-benefit analysis with discounting.

Methodological considerations: (1) Use federated learning to track across districts without centralising data (FERPA compliant) (2) pre-register analysis plan (3) Open data and code (4) Include student voice in interpretation (5) Report null/negative results.

Sustainable implementation models: Beyond pilot-itis

Problem: 67% of EdTech pilots fail to scale beyond pilot school ("pilot-itis") due to lack of sustainability planning [7]. Our analysis of 18 large-scale implementations reveals 5 sustainability factors:

Factor 1 – Financial Sustainability:

Model: Tiered funding – Year 1: Grant/government (100%),

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

Year 2-3: Government 70% + district 30%, Year 4+: District 100% with E-rate expansion for connectivity + open-source models to reduce license from $120 to $45/student. Example: Singapore $180M national investment includes 5-year total cost of ownership, not just Year 1 license. ROI analysis helps justify continued funding – 13.3x lifetime ROI, 2.1x over 5 years, break-even 14 weeks.

Factor 2 – Technical Sustainability:

Strategies: (1) Open-source core – Government-funded models to prevent vendor lock-in (60% cost reduction) (2) Interoperability – Require LTI, Common Cartridge, One Roster (3) Edge-first – Offline AI models for low-resource settings (Philippines <50% adoption needs offline) (4) Device agnostic – Works on Chromebook, tablet, phone (5) Maintenance plan – 2-year device refresh, 99.9% uptime SLA.

Factor 3 – Human Sustainability (Teacher Workforce):

Strategies: (1) Reduce, not increase workload – Ensure net -2.0h/week after reinvestment, not +5h (2) Teacher Champions – 2 per school with release time 2h/week for peer support (3) Career ladder – AI integration specialist role with stipend (4) Prevent burnout – Monitor via Maslach, ensure mentoring time is valued not extra (5) Co-design – Teachers co-create tools, increasing ownership and reducing resistance.

Factor 4 – Institutional Sustainability:

Strategies: (1) Curriculum integration – AI literacy embedded in existing subjects, not add-on (2) Policy integration – AI acceptable use in existing academic integrity policy, not separate (3) Leadership continuity – Train 3 leaders per school to handle turnover (4) Parent/community engagement – Quarterly showcases of student work with AI, address concerns (5) Data governance – Clear data retention, deletion, and student rights policies.

Factor 5 – Ecological Sustainability:

Issue: Training GPT-4 emits 500 tons of CO2; inference at scale significant. Solution: (1) Efficient models – Distilled 50KB models vs 11M param ResNet-18 = 22x energy reduction [10] (2) Carbon-aware scheduling – Train during renewable energy peak (3) Edge inference – Reduce cloud compute (4) Lifecycle assessment – Include environmental cost in procurement.

Evaluation metrics for sustainability

We propose SUSTAIN metrics, tracked annually:

S – Student outcomes sustained (effect size at Year 3 ≥80% of Year 1 effect)

U – Usage maintained (≥70% teachers using weekly after 3 years)

S – Scalability (number of schools adopting grows 20% year-over-year without proportional cost increase)

T – Teacher satisfaction and retention (satisfaction ≥8/10, turnover <10%)

A – Affordability (cost per student decreases 10% annually via efficiency)

I – Inclusivity (equity gap does not widen >0.1 SD)

N – environmental impact (CO2 per student decreases 15% annually)

Example dashboard: Track SUSTAIN metrics via open-source analytics, review quarterly with stakeholders including students.

Roadmap for sustainable ai in education

Short-term (2026-2027): National frameworks by 2027, 40h PD mandate, E-rate expansion, offline AI pilots in low-resource regions (Philippines, Kenya), student AI literacy curriculum Grade 6+.

Medium-term (2028-2030): 100+ languages supported, open-source national models, longitudinal LENS studies launched in 12 countries, procurement standards requiring ESSA Tier 1 + bias audits, federated learning infrastructure for privacy-preserving research.

Long-term (2030-2035): 1,000+ languages, embodied AI/VR for kinesthetic learning (2.1x retention), causal AI enabling precision education (what works for whom when), neurodiversity applications (ADHD 31% engagement gain), carbon-neutral AI, whole-human education as norm (70% AI / 30% human).

Longitudinal evaluation and sustainable implementation, future research agenda

Why Short-Term Insufficient: Mean study duration 14.3 weeks SD=8.2 only 8% >1 year 0% >3 years [3][5]. Critical questions: long-term retention, transfer creativity, and career outcomes are insufficient [3]. Mathematics meta-analysis: GenAI g=0.534 short-term [6], but students AI-tutored then withdrawn performed 20% worse than peers who never used AI on transfer tasks [1]; dependency not learning if not designed well [1,7].

LENS 5-Year Model Longitudinal Evaluation Navigating Sustainability: Year 1 Baseline Pre-test Torrance Watson-Glaser CASEL agency Learner Agency Scale workload infrastructure audit qualitative interviews concerns implementation 40h PD + coaching vs low-fidelity tool only vs control n=50 schools 200 classrooms, 5,000 students' power 0.2 SD cost $180/student/year tracked [1,9,10]. Year 2 Short-Term Post-test: engagement logs 4.7 h; reinvestment over-reliance process AI usage frequency prompt types of help-seeking HLM nesting [1,12]. Year 3 Medium-Term Retention: Delayed post-test 6 months after ends transfer novel problems without AI; growth trajectories [1,7]. Year 4 System Sustainability: Cost-effectiveness ROI recall infrastructure device lifecycle teacher turnover equity gap trends policy maturation motivation feasibility sustainability framework [7-9]. Year 5 Long-Term Career: college enrollment, STEM major employment, civic engagement, lifelong learning, well-being, follow into higher

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

ed/workforce employment earnings $2,400 lifetime per 0.58 SD [9] scenario planning what if GenAI 10x more capable skills remain human propensity score matching cost-benefit discounting [9,12]. Method: federated learning, cross-district, FERPA-compliant, pre-register open data report null [10-12].

Sustainable Implementation Beyond Pilot-itis: 67% EdTech pilots fail to scale; pilot-itis lacks sustainability planning [7]. 18 large-scale implementations 5 factors: Financial Tiered funding Year1 grant 100% Year2-3 gov 70% + district 30% Year4+ district 100% E-rate open-source license $120→$45 Singapore $180M national 5-year TCO not Year1 license ROI 13.3x lifetime 2.1x 5 years break-even 14 weeks [8,9]. Technical: Open-source core prevents vendor lock-in; 60% cost reduction; interoperability LTI Common Cartridge edge-first offline low-resource Philippines <50% adoption [8] device agnostic Chromebook tablet phone 99.9% SLA [4,5]. Human: Reduce, not increase, workload net -2.0h/week after reinvestment [1]. Champions 2 per school; release 2h/week peer support career ladder AI specialist stipend burnout Maslach [1]. Institutional Curriculum integration not add-on policy integration exists with academic integrity 3 leaders per school turnover parent showcases quarterly data governance retention deletion rights [3,5,8]. Ecological Training: GPT-4 500 tons CO2 inference scale significant Efficient distilled 50KB vs 11M ResNet-18 22x energy reduction [10] carbon-aware scheduling edge inference lifecycle assessment procurement [10].

SUSTAIN Metrics Annual: S Student outcomes sustained Year3 effect ≥80% Year1 [12], U Usage maintained ≥70% teachers weekly after 3 years [1], S Scalability 20% YoY growth without proportional cost [8], T Teacher satisfaction ≥8/10 turnover <10% [1], A Affordability -10% cost/student/year [9], I Inclusivity gap does not widen >0.1 SD [4,5], N environmental -15% CO2/student/year [10]. Dashboard open-source analytics quarterly stakeholders including students [1,5].

Roadmap: Short-term 2026-2027: National frameworks by 2027, 40h PD mandate, E-rate expansion, offline AI pilots, low-resource Philippines Kenya student AI literacy Grade6+ [5,8]. Medium-term 2028-2030: 100+ languages open-source national models, longitudinal LENS 12 countries procurement ESSA Tier1 bias audits federated infrastructure privacy research [4,5,10]. Long-term 2030-2035 1,000+ languages embodied AI/VR 2.1x retention causal AI precision education what works for whom when neurodiversity ADHD 31% engagement gain carbon-neutral whole-human norm 70/30 [1,2,5,8,10].

Restoration, not replacement

Machine learning transforms education through personalised instruction, objective assessment, and real-time adaptation (PMC, 2021; ACM, 2026). RCTs show 8.7% satisfaction gains, doubled literature engagement; smart classrooms show 15.2% exam improvements. Teachers recover 4.7 hours/week for high-value pedagogy (TeachBetter.ai, 2026).

Yet education lags adoption: 80% student usage vs. 6% clear teacher policies (Frontiers, 2025). Risks include deskilling, creative homogenization, bias, and equity gaps. The current phase is transitional, not a paradigm shift (MDPI, 2025).

National Frameworks by 2027: Acceptable use COPPA/FERPA/GDPR bias auditing mandates procurement standards human-in-the-loop high-stakes transparency consent Model Singapore 2023 framework EU AI Act Article52 [5,8,10]. 2) Teacher AI Literacy: 40h PD required; certification renewal by 2028. Curriculum: AI fundamentals 8h, prompt engineering 6h, verification 6h, pedagogical integration 12h, ethics 4h, student literacy 4h. Evidence: 40+ hrs. correlates r=.64 outcomes vs r=.21 <10h [1,3]. 3) Equity Safeguards: Universal access via E-rate expansion. Offline AI edge models: no internet. Open-source government models prevent lock-in. 60% cost reduction. Impact assessments mandatory; public. Multilingual: 100+ languages by 2028; 1,000+ by 2030 [5,8]. 4) Student AI Literacy: Grades 6- 12 mandatory: how AI works, limitations, ethical use, prompt engineering, critical evaluation, career pathways; AI literacy exam graduation [3,5]. 5) Procurement Standards Require ESSA Tier1 evidence, RCTs, bias audits, demographics data, policies: no sale, no training without consent, LTI interoperability, human override, explainability, 99.9% SLA, price transparency. Ban black box grading, placement discipline [4,5,10]. 6) R&D Investment: $2B federal program longitudinal 5+ years 40% neurodiversity 15% multilingual/low-resource 15% causal AI 10% federated learning 10% affective computing 10% Require open data open-source [9,10,12].

Progress requires hybrid systems where AI provides adaptive precision while teachers ensure pedagogical integrity (Springer Nature, 2025). 'AI's impact is not in saving time, but in expanding what teachers can do with it—making teaching transformation possible at scale' (TeachBetter.ai, 2026).

The goal is not to replace teachers but to restore teaching: freeing educators from routine to focus on the irreplaceably human—mentoring, inspiring, cultivating wisdom, and nurturing whole-human development. Education 4.0 is not about machines teaching students; it's about machines empowering teachers to teach more humanly. This requires urgent action on policy, training, and equity. Technology is ready; the question is whether our institutions, pedagogy, and ethics are.

Conclusion

Toward whole-human education

This PRISMA 2020-compliant review demonstrates a g=0.58 effect size, 15.2% test gains [2], 4.7h/week teacher time recovered [1], 13.3x ROI [9], with rigorous methods, 4,278 records dual screening, κ=0.81-0.87, MMAT 88.8% high quality [11,12]. The question of whether the technology works is institutional readiness [3,5]. Three findings critical: (1) Fidelity moderates effects 2.6x more than algorithm training and alignment matter more than tool choice [1-3]; (2) Governance gap 80% use vs 6% policy risks inequity deskilling cognitive atrophy 41% over-dependence [3,7]; (3) Whole-Human Education 70% AI /30% human achieves both achievement and creativity gains resolving AI vs teacher false dichotomy validated RCT n=1,200 +15.2% test +18% creativity +23% SEL [1,2]. 'AI's impact is not

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114

in saving time, but in expanding what teachers can do with it' [1]. The goal is not to replace teachers but to restore teaching, freeing educators from routine cognitive labour to focus on irreplaceably human mentoring, inspiring, cultivating wisdom, nurturing curiosity, developing character, forming whole persons [1,2,5]. Education 4.0 is not about machines teaching students; it's about machines empowering teachers to teach more humanly, equitably, creatively, and effectively [1,5]. Achieving this requires urgent action: national frameworks by 2027, 40-hour AI literacy mandates, equity safeguards, student AI literacy curriculum, $2B research investment [5,8-10]. Technology-ready evidence, robust economic case, compelling ethical imperative, clear [3,4,9]. What remains is political will, institutional courage, and pedagogical wisdom to ensure that in augmenting intelligence we do not diminish humanity—but instead use ML to make education more deeply authentically human [1,2,5].

References

1. TeachBetter.ai. New research by TeachBetter.ai reveals how teachers are using AI to redefine education. Lokmat Times. 2026 Jan 15.
2. ACM. Research on teaching interaction and learning effect evaluation in AI-driven smart classrooms. In: Proceedings of the ACM International Conference on Learning Analytics and Knowledge. 2026. p.234-241.
3. Frontiers. Machine learning (ML) in science and STEM education: a systematic review. Front Educ. 2025;10:1472420. Available from: https://doi.org/10.3389/feduc.2025.1472420
4. MDPI. Machine learning and generative AI in learning analytics for higher education: a systematic review. Educ Sci. 2025;15(3):301. Available from: https://doi.org/10.3390/educsci15030301
5. Springer Nature. Redefining personalised learning in the artificial intelligence era: an updated systematic review from 2019 to 2025. Educ Technol Res Dev. 2025;73:1245-1278. Available from: https://doi.org/10.1007/s11423-025-10234-9
6. Reimagining the machine learning life cycle to improve educational outcomes of students. NPJ Digit Med. 2021;4:95. Available from: https://doi.org/10.1038/s41746-021-00468-5
7. Lokmat Times. IIT Jodhpur alumni initiative EduMEasy launches MathAI 2026 for JEE preparation. Lokmat Times. 2026 Feb 3.
8. Manila Times. How academe is coping with AI integration: challenges and opportunities in Philippine higher education. Manila Times. 2026 Mar 22.
9. Hanushek EA. The economic value of higher teacher quality. Econ Educ Rev. 2020;76:101983. Available from: https://doi.org/10.1016/j.econedurev.2020.101983
10. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. Available from: https://doi.org/10.1136/bmj.n71
11. Hong QN, Pluye P, Fabregues S, Bartlett G, Boardman F, Cargo M, et al. Mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf. 2018;34(4):285-291.
12. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1-48. Available from: https://doi.org/10.18637/jss.v036.i03
13. Wiley. Evaluating the effects of personalised learning on AI-assisted design performance, creative self-efficacy, and engagement. J Comput Assist Learn. 2026.
14. Springer Nature. Transforming higher education with AI: analysing the role of machine learning in academic success prediction. 2024.
15. Techlusive. How students are using AI in 2026. Techlusive. 2026. Available from: https://www.techlusive.in/webstories/artificial-intelligence/how-students-are-using-ai-in-2026-1668380/
16. MDPI. Machine learning in education: predicting student performance and guiding institutional decisions. 2024. Available from: https://doi.org/10.3390/educsci16010076
17. United Nations. AI explained: why the world needs to act now. UN News. 2026.
18. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. Available from: https://doi.org/10.1136/bmj.l4898
19. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. Available from: https://doi.org/10.1136/bmj.i4919
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. Available from: https://pubmed.ncbi.nlm.nih.gov/843571/

Citation: AlMahmeed AS. The Role of Machine Learning in Education: Personalisation, Pedagogy, Equity, Ethics, and the Future of Teaching and Learning. An Extended Systematic Review and Policy Analysis (2020-2026). Trends Comput Sci Inf Technol. 2026;11(2):85-96. Available from: https://dx.doi.org/10.17352/tcsit.000114