

Artificial Intelligence in Higher Education From Automation to Responsible Augmentation

An Jiangyu¹, Lim Boon Hooi^{1*}

¹ Faculty of Education and Liberal Arts, Inti International University, 71800 Bandar Baru Nilai, Labu, Seremban, Negeri Sembilan, Malaysia

ARTICLE INFO

Article history:

RECEIVED 27 September 2026

ACCEPTED 10 October 2026

PUBLISHED 12 October 2026

Keywords:

artificial intelligence; higher education;
generative AI; responsible augmentation;
narrative review

ABSTRACT

Artificial intelligence (AI) is often presented as a sequence of increasingly capable tools, yet educational outcomes do not rise automatically with model capability. This narrative review argues that the decisive variable is alignment: AI produces durable educational value when its affordances are aligned with pedagogy, assessment, educator agency, and institutional governance. The review traces the field from automation and learning analytics through adaptive support and generative co-production to an emerging model of responsible augmentation. Recent meta-analyses, controlled studies, systematic reviews, policy documents, and selected empirical studies were critically compared across three interdependent domains: resources and learning environments; teaching and learning; and management, assessment, and evaluation. The synthesis explains why apparently contradictory findings coexist. Structured, curriculum-aligned AI tutoring can improve immediate learning, whereas unconstrained answer generation can raise supported performance while weakening independent performance. Likewise, automated assessment can improve feedback speed but becomes educationally unsafe when opaque scores or unreliable detectors are treated as final decisions. Evidence concerning international and multilingual learners further shows that language, culture, affect, and data representation are not peripheral variables. The review proposes a responsible-augmentation framework in which alignment, agency, equity, and accountability operate across all three domains. It concludes with a research agenda prioritising longitudinal transfer, calibrated human–AI collaboration, subgroup validity, transparent governance, and institution-level implementation studies.

1. Introduction

Artificial intelligence has moved from a specialised educational technology to an everyday layer of university life. Recommendation engines organise learning resources, predictive systems identify students considered at risk, automated scoring tools accelerate feedback, and generative AI can now draft explanations, solve problems, simulate dialogue, and produce multimodal content within seconds. This rapid expansion has encouraged a familiar assumption: if a model becomes more capable, learning must improve. The empirical record does not support such a simple progression.

* Corresponding author: Lim Boon Hooi

Email address: boonhooi.lim@newinti.edu.my

<https://doi.org/10.70702/bdb/ROT15363>

Earlier research on intelligent tutoring systems demonstrated that well-designed adaptive systems can improve learning, particularly when feedback is immediate and instruction is matched to a learner's current state (Ma et al., 2014; Kulik & Fletcher, 2016). More recent meta-analyses continue to report positive average effects for personalised and AI-supported learning, but the magnitude varies markedly by adaptive mechanism, learning domain, duration, comparator, and level of human support (Major et al., 2021; Dong et al., 2025; Wang et al., 2026). Average effectiveness therefore conceals a design problem: the same technical capability can support reflection in one setting and substitute for it in another.

Generative AI makes this tension unusually visible. A carefully engineered tutor can provide hints, ask diagnostic questions, and withhold direct answers. A general-purpose chatbot can instead deliver a polished solution before the learner has represented the problem. Both systems may use similar underlying models, yet they create different cognitive demands. In a controlled physics study, a structured AI tutor produced larger immediate learning gains in less time than an active-learning class, although the intervention was brief and tightly designed (Kestin et al., 2025). By contrast, a field experiment found that unrestricted access to generative AI improved performance during practice but reduced subsequent unaided performance; guardrails substantially mitigated that cost (Bastani et al., 2025). The contradiction is only apparent. What differs is not simply model quality, but the instructional architecture surrounding the model.

This review develops the thesis that **AI's educational value depends less on raw model capability than on alignment across pedagogy, assessment, educator agency, and institutional governance**. Alignment means that the system's role, the learner's task, the target capability, the assessment evidence, and the institution's responsibilities reinforce rather than undermine one another. The argument extends the three concerns that have consistently organised educational reform—resources and environments, teaching and learning, and management and evaluation—by treating them as interdependent layers rather than separate application categories. A resource platform changes what students can access; that change reshapes teaching practices; those practices in turn require different forms of assessment and governance.

The review has four aims. First, it reconstructs the development of AI in education as a shift from automation to responsible augmentation. Second, it synthesises evidence across the three layers while identifying recurrent contradictions and their likely methodological or contextual sources. Third, it foregrounds international and multilingual learners, for whom linguistic bias, cultural fit, and affective engagement can alter both benefits and risks. Fourth, it proposes testable research priorities that move the field beyond short demonstrations and self-reported acceptance. The resulting account is not a catalogue of tools. It is an explanation of when AI strengthens education, when it merely increases task completion, and when it quietly erodes the very capabilities that education is intended to develop.

2. Materials and Methods

2.1 Review design and scope

A narrative review design was selected because the central question is explanatory rather than enumerative: why do studies of educational AI reach different conclusions, and what organising framework can reconcile them? Narrative synthesis is appropriate for a heterogeneous literature that includes meta-analyses, randomised and quasi-experimental studies, systematic and scoping reviews, conceptual frameworks, and policy guidance. The review nevertheless used a transparent search and appraisal process to reduce selective interpretation.

The principal time window was January 2019 to 7 October 2026, capturing the expansion of learning analytics, adaptive systems, and generative AI. Seminal pre-2019 studies were retained when they established a construct or evidence baseline, including meta-analyses of intelligent

tutoring systems and foundational work on learning-analytics ethics. The educational focus was higher education, with selected school-based or workplace studies included when they offered unusually strong causal evidence or illuminated a mechanism directly relevant to universities. This boundary is important: evidence was transferred across settings only at the level of mechanism, not assumed to generalise automatically.

2.2 Search and selection strategy

Candidate literature was located through cross-disciplinary academic discovery services and verified against publisher pages, Crossref, PubMed, or OpenAlex metadata. Search combinations joined four concept groups: (1) “artificial intelligence,” “generative AI,” “large language model,” “adaptive learning,” “intelligent tutoring,” or “learning analytics”; (2) “higher education,” “university,” “international student,” “EFL,” or “multilingual learner”; (3) “learning outcome,” “critical thinking,” “assessment,” “feedback,” “engagement,” or “well-being”; and (4) “ethics,” “privacy,” “bias,” “governance,” or “policy.” Backward and forward citation tracing was used to connect recent studies with established theoretical and empirical foundations.

Priority was given to peer-reviewed meta-analyses, systematic reviews, controlled studies, multi-institutional research, and policy documents from public or intergovernmental bodies. Studies were retained when they contributed evidence about at least one of the three layers or explained a boundary condition. Purely promotional accounts, unverified commentary, tool descriptions without an educational outcome, and sources whose bibliographic identity could not be confirmed were excluded. Because the purpose was critical synthesis rather than exhaustive coverage, no claim of PRISMA-style completeness is made and no new pooled effect was calculated.

2.3 Critical comparison

Each source was read against six questions: What educational capability was targeted? What role did AI perform? What did the comparison condition actually provide? Was the outcome supported performance, independent performance, perception, or durable learning? How long did the intervention and follow-up last? Which learners and institutions were represented? These questions prevented unlike outcomes from being treated as interchangeable. For example, satisfaction with a chatbot was not treated as evidence of knowledge transfer, and the accuracy of an automated score was not treated as evidence that its use was fair.

The synthesis was organised in two passes. The first mapped evidence onto resources and environments; teaching and learning; and management, assessment, and evaluation. The second compared claims across designs to identify convergence, conflict, and missing evidence. Greater interpretive weight was assigned to converging results across designs and to causal studies with meaningful comparison conditions. Policy recommendations were used to clarify institutional obligations, not as substitutes for outcome evidence. This approach produces a reasoned account of the literature while preserving the limitations inherent in a narrative review.

3. Results and Discussion

3.1 The storyline from automation to responsible augmentation

The field’s development is best understood as four overlapping stages rather than a clean sequence. The first stage, **automation and analytics**, used classification, prediction, and rule-based feedback to make existing processes faster. Early-warning systems, automated scoring, and administrative chatbots belong largely to this logic. Their value was operational efficiency, but their central risks—opaque inference, privacy loss, and the conversion of probabilistic predictions into labels—were already evident in learning-analytics scholarship (Slade & Prinsloo, 2013; Pardo & Siemens, 2014). Students’ willingness to share data also depended on purpose, control, and perceived benefit, not simply on technical security (Ifenthaler & Schumacher, 2016).

The second stage, **adaptive support**, shifted attention from institutional prediction to learner-responsive instruction. Intelligent tutors and adaptive courseware selected tasks, feedback, or pathways according to inferred knowledge states. This stage established a crucial principle: adaptation is educationally valuable only when the variables being adapted correspond to a defensible model of learning. Meta-analytic findings are positive overall, yet systems that merely personalise pace should not be equated with systems that diagnose misconceptions or regulate challenge (Major et al., 2021; Wang et al., 2026).

The third stage, **generative co-production**, expanded the system's role from selection to creation. Students and teachers can now co-produce explanations, examples, code, cases, feedback, and assessments. Reviews published after the release of widely accessible large language models describe opportunities for tutoring, accessibility, creativity, and workload reduction alongside hallucination, bias, academic-integrity, and overreliance risks (Kasneci et al., 2023; Lo, 2023; Tlili et al., 2023). The distinctive challenge is epistemic: fluent output can look authoritative while remaining probabilistic, incomplete, or fabricated.

The emerging fourth stage is **responsible augmentation**. Here AI is neither a replacement teacher nor a neutral tool. It is a fallible participant in a deliberately designed activity system. The design question changes from "Can AI perform this task?" to "Which parts should AI perform, which judgements must remain human, and what evidence will show that the partnership improves learning?" Human–AI complementarity, professional agency, and accountable governance become design requirements rather than ethical add-ons (Molenaar, 2022; Holmes et al., 2022). Figure 1 summarises this trajectory and the unresolved question at each stage.

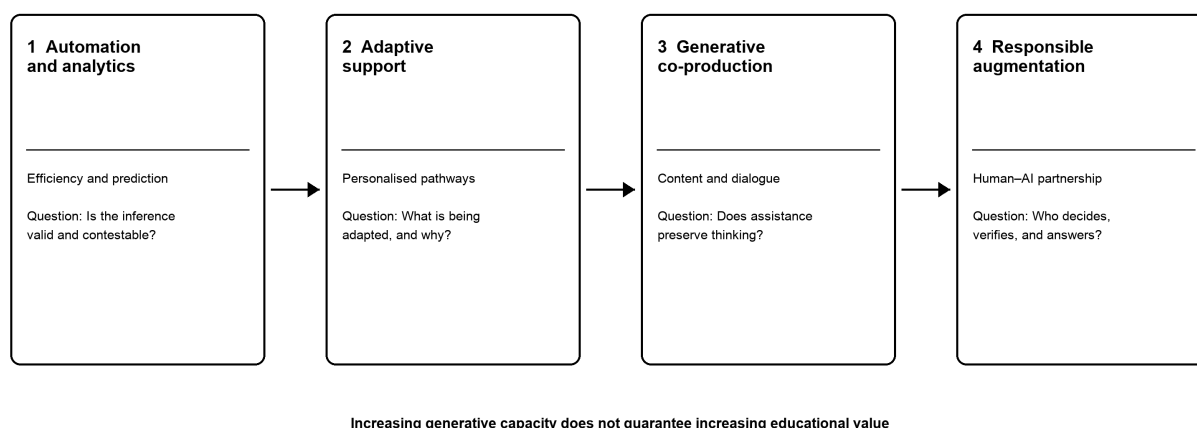


Figure 1. From automation to responsible augmentation: changing questions for educational AI

This periodisation also explains why institutional policy frequently lags practice. Universities often respond to a generative disruption with rules designed for an automation problem—for example, attempting to detect prohibited text rather than redesigning the production and verification of knowledge. A responsible-augmentation approach instead asks how tasks, evidence, disclosure, and supervision should change when generation is widely available.

3.2 Resources and learning environments

AI can widen access to educational resources by translating, summarising, adapting reading level, generating examples, and providing assistance outside scheduled teaching hours. For international and multilingual students, these affordances may reduce the linguistic cost of entry into unfamiliar disciplinary discourse. Adaptive systems can also route learners to materials that match current performance. Reviews of AI in higher education consistently identify profiling,

prediction, adaptive systems, and intelligent tutoring as major application clusters (Zawacki-Richter et al., 2019; Crompton & Burke, 2023; Bond et al., 2024).

Access, however, is not the same as inclusion. A system can make resources technically available while reproducing linguistic or cultural assumptions in examples, feedback, or difficulty estimates. Large language models inherit uneven representation from training data and may generate stereotypes or privilege dominant forms of English. The problem is compounded when institutional procurement treats a general-purpose product as a ready-made learning environment. The system may be convenient yet misaligned with local curricula, disability accommodations, data-protection requirements, or students' device and connectivity constraints.

The strongest evidence for personalised learning therefore supports conditional, not universal, optimism. A meta-analysis of technology-supported personalised learning in low- and middle-income countries reported a modest positive overall effect and a larger effect for systems adapting to learners' current level, but interventions and contexts were diverse (Major et al., 2021). A later meta-analysis of AI-assisted learning likewise reported a large pooled effect across included studies, while substantial variation indicated that design and context moderated results (Dong et al., 2025). The most recent synthesis of adaptive learning found medium positive effects across cognitive, affective, and behavioural outcomes, again with meaningful moderation (Wang et al., 2026). These estimates should not be read as a single "AI effect." They aggregate distinct systems, populations, subjects, and comparison conditions.

A better interpretation is that AI can reduce the distance between available content and a learner's present needs, but only when diagnosis is valid and the adaptation preserves productive difficulty. Excessive simplification may improve completion while weakening disciplinary reading; constant hints may improve local success while preventing self-explanation. Resource design should consequently include a progression from support to independence, accessible alternatives, and a visible route for learners to challenge or correct system output.

3.3 Teaching and learning

AI's most defensible pedagogical role is not answer production but the orchestration of practice, feedback, explanation, and reflection. This principle links the earlier intelligent-tutoring literature with contemporary generative systems. Meta-analyses of intelligent tutors found meaningful average gains, particularly where systems modelled learners' knowledge and provided targeted feedback (Ma et al., 2014; Kulik & Fletcher, 2016). Yet the label "tutor" is now applied to systems ranging from carefully sequenced curricula to open chat interfaces. Their pedagogical functions are not equivalent.

The contrast between two recent controlled studies is instructive. Kestin and colleagues embedded an AI tutor within an introductory physics sequence and constrained it with pedagogical instructions, curricular content, and structured interaction. Students learned more on immediate post-tests and spent less time than peers in an active-learning class (Kestin et al., 2025). The finding is important, but its scope is narrow: two lessons, a single course context, immediate outcomes, and a bespoke tutor. It demonstrates the potential of aligned design, not the superiority of AI over classroom teaching in general.

Bastani and colleagues reached a different result because they measured a different outcome under a different support architecture. Students given unrestricted generative AI performed better while the tool was available but worse when assistance was removed. A tutor-like version with guardrails reduced that penalty (Bastani et al., 2025). Supported performance and independent learning had diverged. Together, the studies suggest a general mechanism: AI helps when it increases cognitively relevant activity—diagnosis, retrieval, explanation, comparison, and revision—and harms when it compresses that activity into answer acquisition.

This mechanism also clarifies concerns about critical thinking. In a survey of knowledge workers, greater confidence in generative AI was associated with less reported critical-thinking effort, whereas confidence in one's own ability was associated with more effort (Lee et al., 2025). The study did not test students or directly measure cognition, so it cannot establish learning loss. It nevertheless identifies a plausible calibration problem: when fluent output inflates trust, users shift effort from constructing and checking claims to accepting them. Teaching therefore needs to make verification observable. Students can be required to identify uncertainty, compare model responses with primary sources, document rejected suggestions, or defend revisions orally.

Teacher agency is central to this design. Reviews of educator–AI collaboration show that teachers require more than operational familiarity; they need pedagogical and ethical knowledge to decide when AI should diagnose, recommend, generate, or remain absent (Celik et al., 2022). AI literacy likewise includes understanding what systems can and cannot know, how data and models shape outputs, and how to evaluate consequences (Long & Magerko, 2020; Ng et al., 2021). Professional development should therefore begin with learning problems and evidence of student thinking, not a tour of popular tools.

Student perceptions remain relevant but insufficient. Learners often value generative AI for explanation, brainstorming, language support, and efficiency while expressing concern about accuracy, privacy, and dependency (Chan & Hu, 2023). These perceptions can explain adoption and disengagement, but they do not establish effectiveness. Table 1 distinguishes what current evidence reasonably supports from what remains conditional in each educational layer.

Table 1. Evidence and boundary conditions across educational functions

Domain	What evidence supports	Boundary or unresolved risk	Design implication
Resources and environments	Adaptive systems can improve access, pacing, and match to current level; translation and explanation can reduce entry barriers.	Diagnosis may be wrong; dominant-language and cultural assumptions can narrow inclusion; simplification can remove productive difficulty.	Offer accessible alternatives, expose uncertainty, and plan the withdrawal of support.
Teaching and learning	Structured tutoring, targeted feedback, and guided practice can improve immediate learning and efficiency.	Unrestricted answer generation can produce dependency; satisfaction and supported performance do not establish transfer.	Use prompts and task sequences that require retrieval, explanation, verification, and revision.
Assessment and evaluation	Automated scoring and feedback can increase consistency and response speed in bounded tasks.	Construct under-representation, opacity, and biased AI-text detection threaten validity and fairness.	Triangulate products with process evidence, oral defence, and human review; never treat detector output as proof.
Management and governance	Prediction and automation can focus staff attention and reduce routine workload.	False labels, surveillance, privacy loss, and unclear responsibility can offset operational gains.	Define decision rights, appeals, audit indicators, and a clear owner for every consequential use.

Note. The table distinguishes an educationally plausible mechanism from a product-level claim.

3.4 Management assessment and evaluation

Administrative uses of AI are attractive because they promise scale: triaging inquiries, scheduling, identifying students at risk, processing applications, or producing dashboards. These functions can release staff time, but efficiency gains become educational gains only if the saved capacity is redirected toward meaningful support. Predictive accuracy alone is not an adequate criterion. A risk score can change how staff interpret a student, creating self-fulfilling expectations; a false positive can trigger unnecessary surveillance, while a false negative can hide need.

Institutions must define who can act on a prediction, what evidence accompanies it, and how students can contest it.

Assessment presents the most consequential alignment problem. Automated essay scoring research shows that machines can produce consistent scores in constrained contexts, yet validity depends on the construct represented by training data and features (Ramesh & Sanampudi, 2022). Generative AI increases the need to distinguish product from process. If an assignment can be completed through unexamined generation, a polished final text provides weak evidence of the learner's reasoning. The appropriate response is not simply to prohibit tools. Assessment can distribute evidence across proposals, source trails, drafts, feedback decisions, oral defence, in-class application, and reflective accounts of tool use.

AI-text detection is a particularly weak foundation for high-stakes judgement. Independent testing has found substantial variation and evasion vulnerability across detection tools (Weber-Wulff et al., 2023). More seriously, detectors have misclassified writing by non-native English authors at very high rates, showing how linguistic conformity can be mistaken for machine generation (Liang et al., 2023). This is not a minor measurement error. For international students, it can transform language background into suspicion and place the burden of proof on the learner. Detection output should therefore never be treated as dispositive evidence of misconduct.

Assessment redesign should be guided by three questions. First, what human capability is the assessment intended to infer? Second, which forms of AI assistance are consistent with that capability? Third, what combination of artefacts and performance can support the inference? Bearman and colleagues describe assessment in a digital world as a question of how trustworthy judgements can be made amid changing practices (Bearman et al., 2023). The practical implication is that AI use should be specified at task level: prohibited where unaided fluency is the construct; permitted with disclosure where tool-mediated performance is authentic; and required where critical use of AI is itself an intended outcome.

Institutional policies are gradually moving from blanket prohibition toward capability, transparency, and risk-based governance. Frameworks for university policy emphasise learning, fairness, academic integrity, privacy, and human oversight (Chan, 2023). International guidance similarly recommends human-centred use, age-appropriate safeguards, data protection, and validation rather than uncritical deployment (UNESCO, 2023; European Commission, 2026). Policy must still be translated into workload, procurement, training, appeals, and audit procedures; otherwise principles remain aspirational.

3.5 Why apparently contradictory findings coexist

Disagreement in the literature frequently arises because studies use the same term—"AI-assisted learning"—for different interventions and outcomes. Five distinctions are especially important.

First, **access to an answer is not equivalent to learning**. Practice scores obtained with AI measure joint performance; post-tests without AI measure what the learner can independently retrieve or apply. Second, **personalisation differs in depth**. Adjusting pace, selecting a difficulty level, diagnosing a misconception, and adapting motivational support are distinct mechanisms. Third, **comparison conditions vary**. AI may look powerful against passive materials but offer little advantage over well-supported active teaching. Fourth, **time horizons matter**. Immediate post-tests capture short-term performance; delayed transfer tests reveal whether knowledge is durable and flexible. Fifth, **context moderates both benefit and risk**. Language proficiency, subject structure, prior knowledge, teacher mediation, and institutional resources influence how learners use support.

Table 2 applies these distinctions to prominent tensions. The purpose is not to select a winning study, but to show how methodological choices locate different parts of the same causal system.

Table 2. Why apparently contradictory findings coexist

Empirical contrast	Mechanism that differs	Methodological explanation	Interpretation
Structured physics tutor shows gains; unrestricted mathematics chatbot harms unaided performance	Scaffolding and answer withholding versus rapid answer provision	Different populations, subjects, duration, tutor design, and outcome timing	The instructional harness, not the language model alone, is the active treatment.
Meta-analyses report positive means; individual studies vary widely	Depth of adaptation and teacher mediation	Heterogeneous systems, comparators, domains, and study quality are pooled	Average effects indicate potential, not a universal treatment effect.
Students report usefulness; evidence of durable learning is limited	Convenience and affect versus retrieval and transfer	Self-report and immediate task success measure different constructs from delayed performance	Adoption data should not be used as learning evidence.
Detection tools promise integrity; audits show evasion and language bias	Statistical text regularity is treated as authorship	Base rates, model change, paraphrasing, and multilingual writing alter error rates	Detector scores are unsuitable as stand-alone evidence in high-stakes decisions.

Note. The table distinguishes an educationally plausible mechanism from a product-level claim.

The synthesis points toward a more precise explanatory model. Outcomes are produced by an interaction among model affordances, instructional constraints, learner characteristics, and the evidence demanded by assessment. Increasing affordance without strengthening constraints can make tasks easier while making learning less visible. Conversely, increasing control without preserving agency can turn adaptive support into surveillance or rigid sequencing. Responsible design balances assistance with challenge and makes that balance revisable through evidence.

3.6 A responsible augmentation framework

Figure 2 presents an integrated framework. The three horizontal layers represent the educational functions inherited from earlier reform agendas: resources and environments; pedagogy and learning; and management, assessment, and governance. Four vertical conditions—alignment, agency, equity, and accountability—must hold across all layers. A weakness in one layer propagates through the others. For example, culturally narrow resources can distort feedback; distorted feedback can affect assessment; assessment data can then feed institutional predictions, amplifying the original bias.

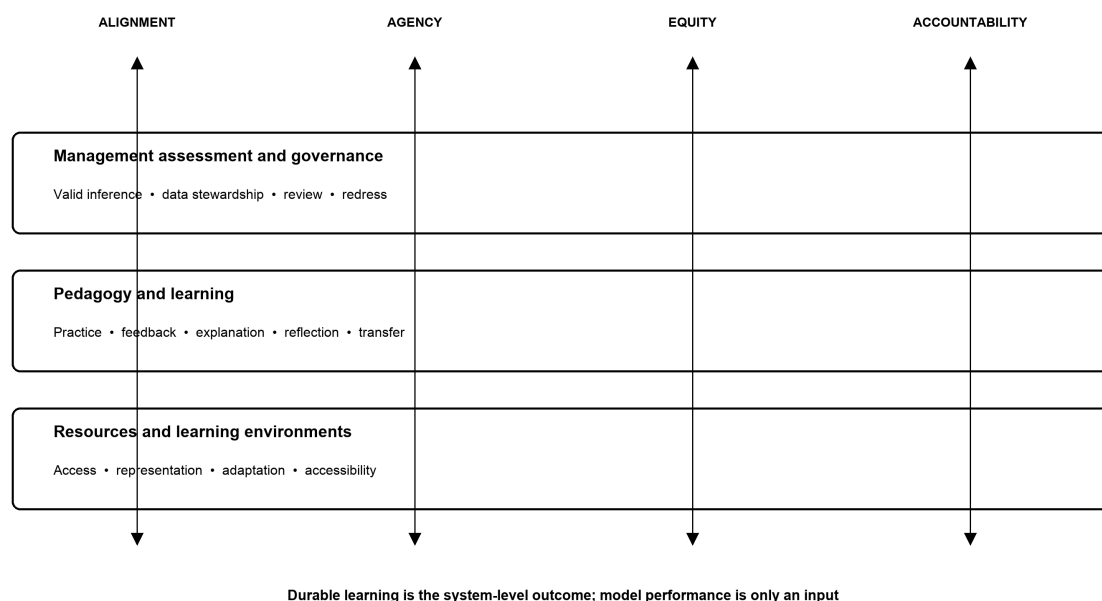


Figure 2. Responsible augmentation framework across three educational layers

Alignment requires a match among the intended capability, AI role, learning activity, and assessment evidence. **Agency** means that teachers and learners can understand, override, question, and meaningfully choose how a system is used. **Equity** requires subgroup validation, accessibility, language-sensitive design, and attention to who bears the costs of error. **Accountability** identifies who is responsible for procurement, data stewardship, monitoring, explanations, and remedies when harm occurs. These conditions convert abstract ethics into design and governance decisions.

The framework also implies a sequence for implementation. Institutions should begin with a defined educational problem, establish a baseline, identify the human judgement that must be preserved, and select the narrowest AI role capable of adding value. A limited pilot should then assess both intended outcomes and counter-metrics such as dependency, unequal error, privacy burden, or staff workload. Scale is justified only when benefits persist beyond tool access and across relevant student groups. This sequence is compatible with calls for an effective digital education ecosystem, but it adds a sharper criterion: ecosystem maturity is demonstrated by the ability to stop, revise, and contest AI use, not merely by the number of systems deployed (OECD, 2023; U.S. Department of Education, 2023).

3.7 International multilingual and EFL learners

International students encounter AI through a distinctive combination of opportunities and risks. On the benefit side, AI can provide low-stakes language rehearsal, tailored examples, immediate feedback, and support outside classroom hours. These functions may lower anxiety and expand willingness to communicate. In a 2026 study of Chinese EFL learners, an AI-supported motor-skill language-learning intervention was associated with stronger gains in willingness to communicate, with enjoyment and grit acting as partial mediators (Long et al., 2026). The result matters because it shows that AI's effect can travel through affective and motivational pathways rather than through content delivery alone. Its specific context and intervention, however, caution against generalising to all language learning.

Related work links curriculum and pedagogical practice with learner competence, emotional well-being, intercultural sensitivity, and classroom engagement, reinforcing the point that technology operates inside a broader instructional ecology (Pan et al., 2026). AI that generates culturally unfamiliar examples or treats one English norm as universally correct may weaken that

ecology. Conversely, tasks that invite students to compare variants, explain culturally situated choices, and use AI output as an object of critique can strengthen intercultural and metalinguistic awareness.

The risks are most acute when language difference is converted into a disciplinary signal. The documented bias of AI-text detectors against non-native English writing demonstrates how an apparently technical solution can reproduce inequity (Liang et al., 2023). Fair practice requires more than publishing an acceptable-use policy. Universities should provide examples of permitted assistance, avoid detector-led accusations, ensure access to comparable tools, protect sensitive learner data, and offer a human review process in which students can present drafts and explain their decisions.

This population also exposes a broader limitation in the literature: the global circulation of AI systems is not matched by globally representative evidence. Much research is conducted in well-resourced institutions, in English, and over short periods. Studies in low- and middle-income contexts indicate positive potential but also substantial contextual dependence (Major et al., 2021). Future research must treat language, culture, connectivity, and institutional capacity as explanatory variables rather than demographic footnotes.

3.8 Research gaps and a testable agenda

The most urgent gap is **durable independent learning**. Many studies measure performance while assistance is available or immediately afterwards. Delayed assessments and transfer tasks are needed to test whether support becomes competence. Designs should compare unrestricted generation, scaffolded tutoring, non-AI active learning, and delayed withdrawal of support. Cognitive-process data—revision histories, explanations, error correction, and help-seeking sequences—can reveal why outcomes differ.

A second gap concerns **calibrated human–AI collaboration**. The field often compares AI with humans when the practical question is how responsibilities should be allocated between them. Experiments should manipulate who initiates feedback, who verifies claims, and when the system reveals an answer. Teacher-facing studies should measure decision quality, workload redistribution, and professional learning, not only intention to adopt.

Third, **subgroup validity and distributional effects** remain underexamined. Aggregate gains can coexist with harm to multilingual students, students with disabilities, novices, or learners with limited access. Researchers should pre-specify subgroup analyses, report error distributions, test measurement equivalence, and involve affected students in design. Equity is not demonstrated by equal access to a system whose errors are unequal.

Fourth, **assessment validity in AI-rich environments** requires longitudinal and programme-level work. Individual assignments can be redesigned, but institutions need evidence about whether portfolios, oral defence, process logs, or authentic performance improve the validity and feasibility of assessment across a curriculum. Studies should report staff time, student burden, appeals, accessibility, and consistency as well as misconduct rates.

Fifth, **governance effectiveness** is largely asserted rather than tested. Policy frameworks identify principles, yet little evidence compares procurement controls, disclosure regimes, audit procedures, or redress mechanisms. Natural experiments across institutions and implementation studies can examine which arrangements reduce harm without suppressing useful innovation. Table 3 translates these gaps into researchable questions and minimum outcomes.

Table 3. Priority research agenda for responsible augmentation

Gap	Testable question	Suitable design	Minimum outcome set
-----	-------------------	-----------------	---------------------

Gap	Testable question	Suitable design	Minimum outcome set
Durable independent learning	Which scaffolds convert assisted performance into unaided transfer?	Multi-arm trials with delayed withdrawal and transfer tests	Immediate, delayed, and unaided performance; error type; time on task
Human–AI role allocation	Which decisions should be initiated, verified, or overridden by teachers and learners?	Factorial experiments plus workflow observation	Decision quality, workload redistribution, agency, and calibration
Subgroup validity	Who benefits, who is misclassified, and under what linguistic or accessibility conditions?	Pre-specified subgroup and measurement-invariance studies	Error distributions, accessibility, differential effects, and appeals
Assessment redesign	Which combinations of product, process, and performance yield trustworthy inference?	Programme-level longitudinal comparison	Validity evidence, staff time, student burden, consistency, and transfer
Governance effectiveness	Which procurement, audit, disclosure, and redress arrangements reduce harm?	Cross-institution implementation studies and natural experiments	Incident rates, response time, contestability, equity, and educational benefit

Note. The table distinguishes an educationally plausible mechanism from a product-level claim.

Finally, the field needs stronger reporting norms. Authors should identify model version, access date, prompting or system instructions, data flows, teacher mediation, permitted student use, comparison condition, and whether outcomes were measured with or without AI. Rapid model change does not make cumulative science impossible, but it makes mechanism-level description essential. Reporting only a product name produces evidence that expires with the interface.

4. Conclusion

The literature does not support a verdict that AI is either inherently transformative or inherently corrosive. It supports a more demanding conclusion: educational effects are produced by the way AI is positioned within a learning system. Structured tutors can focus attention, adapt practice, and expand feedback. Unconstrained generation can improve immediate task performance while weakening independent capability. Automated assessment can accelerate response, but opaque scoring and unreliable detection can damage validity and equity. These are not unrelated successes and failures; they are consequences of alignment or misalignment.

The shift from automation to responsible augmentation therefore changes the unit of evaluation. Models should not be judged only by accuracy, fluency, or adoption. They should be judged by whether learners develop durable capability, teachers retain meaningful professional agency, assessment supports valid inference, and institutions can explain and remedy consequential decisions. For international and multilingual learners, this standard must include language-sensitive validity, cultural fit, and protection from biased detection.

The proposed framework integrates resources and environments, pedagogy and learning, and management, assessment, and governance through four cross-cutting conditions: alignment, agency, equity, and accountability. Its practical message is deliberately conservative in one sense and ambitious in another. Institutions should use the narrowest AI role that solves a genuine educational problem, but evaluate that role across the full system it changes. The next phase of research should move beyond demonstrations of what AI can produce and establish what students can understand, transfer, and responsibly do because of it.

Acknowledgement

The author acknowledges the scholarly communities whose empirical and conceptual work made this synthesis possible.

Funding

No external funding was received for this review.

Conflict of Interest

The author declares no conflict of interest.

Data Availability

No new dataset was generated. All sources analysed are identified in the reference list.

Declaration on the use of generative artificial intelligence and AI-assisted technologies

AI-assisted tools were used during preparation for literature discovery, reference checking, language refinement, and document production. The author remains responsible for verifying the cited sources, evaluating the argument, revising the manuscript, and approving the submitted version.

References

- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, 86, 369–385. <https://doi.org/10.1007/s10734-022-00937-2>
- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, Article 4. <https://doi.org/10.1186/s41239-023-00436-z>
- Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, 66, 616–630. <https://doi.org/10.1007/s11528-022-00715-y>
- Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20, Article 38. <https://doi.org/10.1186/s41239-023-00408-3>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, Article 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20, Article 22. <https://doi.org/10.1186/s41239-023-00392-8>
- Dong, L., Tang, X., & Wang, X. (2025). Examining the effect of artificial intelligence in relation to students' academic achievement in classroom: A meta-analysis. *Computers & Education: Artificial Intelligence*, 8, 100400. <https://doi.org/10.1016/j.caeai.2025.100400>
- European Commission, Directorate-General for Education, Youth, Sport and Culture. (2026). *Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators*. Publications Office of the European Union. <https://doi.org/10.2766/9548952>
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32, 504–526. <https://doi.org/10.1007/s40593-021-00239-1>
- 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>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

- Kestin, G., Miller, K., Klaes, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. <https://doi.org/10.3102/0034654315581420>
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 1121, pp. 1–22). Association for Computing Machinery. <https://doi.org/10.1145/3706598.3713778>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376727>
- Long, J., Yang, Q., Peng, Y., & Pan, L. (2026). Exploring the role of AI-supported motor-skill language learning tools in EFL learners' willingness to communicate: The mediating roles of grit and enjoyment. *Perceptual and Motor Skills*. Advance online publication. <https://doi.org/10.1177/00315125261487804>
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. <https://doi.org/10.1037/a0037123>
- Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology-supported personalised learning in low- and middle-income countries: A meta-analysis. *British Journal of Educational Technology*, 52(5), 1935–1964. <https://doi.org/10.1111/bjet.13116>
- Molenaar, I. (2022). Towards hybrid human–AI learning technologies. *European Journal of Education*, 57(4), 632–645. <https://doi.org/10.1111/ejed.12527>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers & Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Organisation for Economic Co-operation and Development. (2023). *OECD digital education outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing. <https://doi.org/10.1787/c74f03de-en>
- Pan, L., Xu, M., Xiao, Y., & Wang, Z. (2026). EFL curriculum and pedagogical practices as drivers of learner competence, emotional well-being, intercultural sensitivity and classroom engagement. *European Journal of Education*, 61(2), e70689. <https://doi.org/10.1111/ejed.70689>
- 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>

- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529. <https://doi.org/10.1177/0002764213479366>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, Article 15. <https://doi.org/10.1186/s40561-023-00237-x>
- UNESCO. (2023). *Guidance for generative AI in education and research*. <https://doi.org/10.54675/EWZM9535>
- U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. <https://tech.ed.gov/ai-future-of-teaching-and-learning/>
- Wang, J., Tigelaar, D. E. H., Ye, T., & Admiraal, W. (2026). A meta-analysis of moderators of the effects of technology-enhanced adaptive learning on primary and secondary students' learning outcomes. *Journal of Computer Assisted Learning*, 42(1), e70168. <https://doi.org/10.1002/jcal.70168>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>