sustainability

MDPI

Review

Generative AI and Pedagogical Sustainability in Higher Education: A Critical Integrative Review with a European Application

Marcos Komodromos

Department of Management, School of Business, University of Nicosia, 46 Makedonitissas Avenue, P.O. Box 24005, 1700 Nicosia, Cyprus; komodromos.m@unic.ac.cy

Abstract

Generative artificial intelligence (GenAI) promises personalized support, feedback and efficiency in higher education, but raises concerns about cognitive dependency, unequal capability, assessment validity, privacy, environmental cost and commercial dependence. This critical integrative review draws on a purposive, criteria-based selection of 77 sources from Scopus, Web of Science, ERIC and European Union and European Higher Education Area (EHEA) policy portals, classified by type, context and evidence level; key evidence syntheses were appraised with AMSTAR-2 or the JBI scoping review checklist. Meta-analyses report moderate positive average effects on achievement, motivation and higher-order thinking ($g \approx 0.45$–0.80), but heterogeneity is high, very large language-skill effects are probably inflated by small studies, evidence of metacognitive benefit is weak, and durability is untested. We define pedagogical sustainability as the production of durable, transferable learning under economic, social, environmental and institutional conditions that can be maintained over time. We identify five tensions and propose the Sustainable GenAI Pedagogy Framework (SGPF), in which demonstrable learning value is a necessary condition rather than a tradeable pillar. We compare it with eight frameworks and operationalize an Institutional GenAI Maturity Model with indicators, scoring rules and an illustrative application. The framework is general in design and calibrated to European regulation.

Keywords: generative artificial intelligence; pedagogical sustainability; higher education; AI literacy; assessment validity; AI-text detection; EdTech political economy; EU AI Act; Bologna Process; institutional maturity

1. Introduction

The public release of ChatGPT in November 2022 gave students and educators general-purpose generative tools at unprecedented speed. Unlike earlier task-specific educational AI, such as adaptive tutors or automated scorers, generative artificial intelligence (GenAI) generates, summarizes and translates text, writes code, explains concepts and gives feedback through natural-language interaction.

Scholarship has expanded accordingly, but unevenly. A systematic mapping review of 407 publications identified eight discursive themes, led by applications and impacts, ethical implications and risks, user perspectives, institutional and individual adoption, and system performance, and highlighted few experimental studies and limited attention to cultural dimensions [1]. A meta-analysis of 57 studies reported positive average effects on university students' learning outcomes but no significant effect on metacognition and substantial

Check for updates

Academic Editor: Fernando Moreira

Received: 11 October 2026
Revised: 30 October 2026
Accepted: 3 November 2026
Published: 4 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

heterogeneity [2]. Such findings make simple narratives of GenAI either transforming or threatening education hard to sustain.

This review therefore asks: Under what conditions can GenAI produce durable learning without undermining the economic, social, environmental and institutional conditions on which that learning depends?

1.1. Defining Pedagogical Sustainability

The term "pedagogical sustainability" is used in the literature in at least three senses: the durability of learning gains, the sustainability of institutional practice, and the sustainability of education systems. These are distinct constructs, so this review uses one definition throughout:

> **Pedagogical sustainability** is the capacity of a pedagogical practice to produce learning that endures and transfers beyond the immediate task (durable learning value), under economic, social, environmental and institutional conditions that can be maintained over time without eroding learners' independent capability.

The definition has a core and a boundary. The core is durable learning value: retention, transfer and independent capability, not task performance or perceived usefulness. The boundary is the set of four conditions under which that value is produced. Institutional sustainability of practice and system-level sustainability are therefore not alternative meanings but conditions within a single construct.

Two neighbouring concepts are distinguished. Education for sustainable development (ESD) concerns the content and competences of learning about sustainability, as in UNESCO's ESD for 2030 roadmap and the EU GreenComp framework [3,4]. Pedagogical sustainability, by contrast, concerns whether a way of teaching continues to produce learning; it is agnostic about subject matter. Sustainable assessment [5] holds that assessment should equip students to meet their own future learning needs. We adopt it as one mechanism within the durability criterion, not as a synonym for the whole construct.

The definition also explains why demonstrable learning value sits at the center of the framework rather than serving as a fifth pillar. The four pillars describe costs and conditions; learning value is the benefit against which they are weighed. Treating it as a pillar would imply that an application with no learning benefit could be redeemed by being cheap, accessible, green and well governed. Placing it at the center makes it a necessary condition that no pillar can compensate for (Section 13).

1.2. Contribution Relative to Existing Integrative Work

Several recent frameworks already connect two or more of pedagogy, integrity, governance and sustainability [6–9], alongside international guidance and competence frameworks [10–12] and sector maturity models [13]. Our contribution is narrower and more specific; the comparison in Section 13.5 sets it out point by point:

1. A single, bounded definition of pedagogical sustainability that separates durable learning from task performance.
2. A gatekeeping architecture in which learning value is a non-compensable condition, assessed against four pillars, including economic and environmental accounting, which none of the compared frameworks combine.
3. Explicit anchoring to multilevel European regulation and quality assurance (AI Act, GDPR, European quality-assurance standards, EHEA communiqués), while keeping the architecture general.
4. An operational maturity model with indicators, evidence requirements and a scoring rule, applied to an illustrative case.

The review has four aims: (1) to appraise the evidence on GenAI's pedagogical value with explicit attention to evidence quality; (2) to examine its economic, social, environmental and institutional conditions; (3) to identify gaps from a European perspective; and (4) to develop and illustrate a framework and research agenda.

2. Review Method

2.1. Design

This review follows the integrative review methodology of Whittemore and Knafl [14], which combines evidence syntheses, primary studies, and policy texts, and uses the PRISMA 2020 flow structure [15] to describe the selection stages without stage-by-stage record counts. It is not a systematic review of primary studies: its unit of synthesis is mainly secondary evidence (meta-analyses and systematic, mapping and scoping reviews), supplemented by European primary studies and legal and policy documents.

2.2. Information Sources and Search

We identified sources through purposive searches of three bibliographic databases (Scopus, Web of Science Core Collection, and ERIC), using combinations of terms for the technology, setting, and type of evidence, and restricted to publications from January 2023 onwards. Policy and regulatory sources were identified on EUR-Lex, the European Commission's education portal, the EHEA website and the UNESCO digital library, and backward citation searching of the included evidence syntheses identified further sources. Foundational theoretical works (on metacognition, cognitive offloading, the digital divide, the capability approach, the triple bottom line and sustainable assessment) were added purposively and are reported as a separate stratum. Table 1 sets out the search terms in a structured form so that the selection can be replicated and updated. Searches were not logged at each stage, so record counts are not reported (Section 17.1).

Table 1. Structured search terms for source identification, replication and future updating (Scopus syntax).

Block	String
A. Technology	("generative AI" OR "generative artificial intelligence" OR GenAI OR ChatGPT OR "large language model*" OR LLM*)
B. Setting	("higher education" OR universit* OR "tertiary education" OR college*)
C1. Evidence syntheses	("systematic review" OR "meta-analys*" OR "scoping review" OR "mapping review" OR "umbrella review")
C2. European primary evidence	(Europe* OR EU OR "European Union" OR names of EU member states, EEA states, the United Kingdom and other EHEA countries)
C3. Thematic	("academic integrity" OR "AI detect*" OR governance OR "AI literacy" OR "environmental impact" OR carbon OR cost OR "digital divide")
Combination	A AND B AND (C1 OR (C2 AND C3))

Note. The asterisk (*) is a truncation symbol that retrieves all endings of a word stem; for example, universit* retrieves both university and universities.

2.3. Eligibility Criteria

Included: (a) peer-reviewed evidence syntheses (meta-analysis, systematic, mapping or scoping review) on GenAI in higher education; (b) peer-reviewed primary studies conducted in European higher education settings, or, where syntheses are absent or limited, studies addressing assessment validity, AI-text detection, learner agency, AI literacy and the capability divide, feedback design, governance or environmental impact; (c) EU and EHEA legal, policy and guidance documents and international guidance on GenAI in education; (d) English language; (e) published 2023 or later, except foundational theory,

critical EdTech scholarship, legal and policy instruments still in force, and two earlier studies retained as baselines (a pre-generative review of AI in higher education and an analysis of the energy cost of training language models). A small number of supporting sources on related topics (critical thinking, AI-supported personalization, digital competence and sustainability in higher education) were retained for specific claims and are identified in Appendices A and B.

Excluded: (a) studies of pre-generative educational AI only, except the two baseline studies noted above; (b) K-12 settings without a higher-education sample; (c) editorials and non-peer-reviewed preprints, except policy documents (peer-reviewed opinion and perspective articles were retained only as conceptual sources, L4); (d) studies whose GenAI component could not be isolated; (e) duplicate reports of the same dataset.

2.4. Selection Process

Figure 1 sets out the selection process. Because the review follows an integrative rather than a systematic design, records were not counted at each stage, so the figure reports the selection logic and the composition of the final corpus rather than PRISMA record counts. Sources were identified as described in Section 2.2. Candidate sources were screened first on title and abstract and then on full text against the criteria in Section 2.3. Sources were excluded when they addressed pre-generative AI only, had no higher-education sample, were editorials or not peer reviewed, or did not allow the GenAI component to be isolated. The process yielded 77 sources: 18 evidence syntheses (L1 to L3), 15 legal, policy and guidance documents (L4), 17 conceptual and framework papers (L4), 21 primary empirical studies (P) and 6 foundational theoretical works (T), the last added purposively, as described in Section 2.2. Six methodological references are cited separately and are not part of the synthesized corpus. The absence of stage-by-stage record counts limits the reproducibility of the selection, and this is acknowledged as a limitation in Section 17.1.

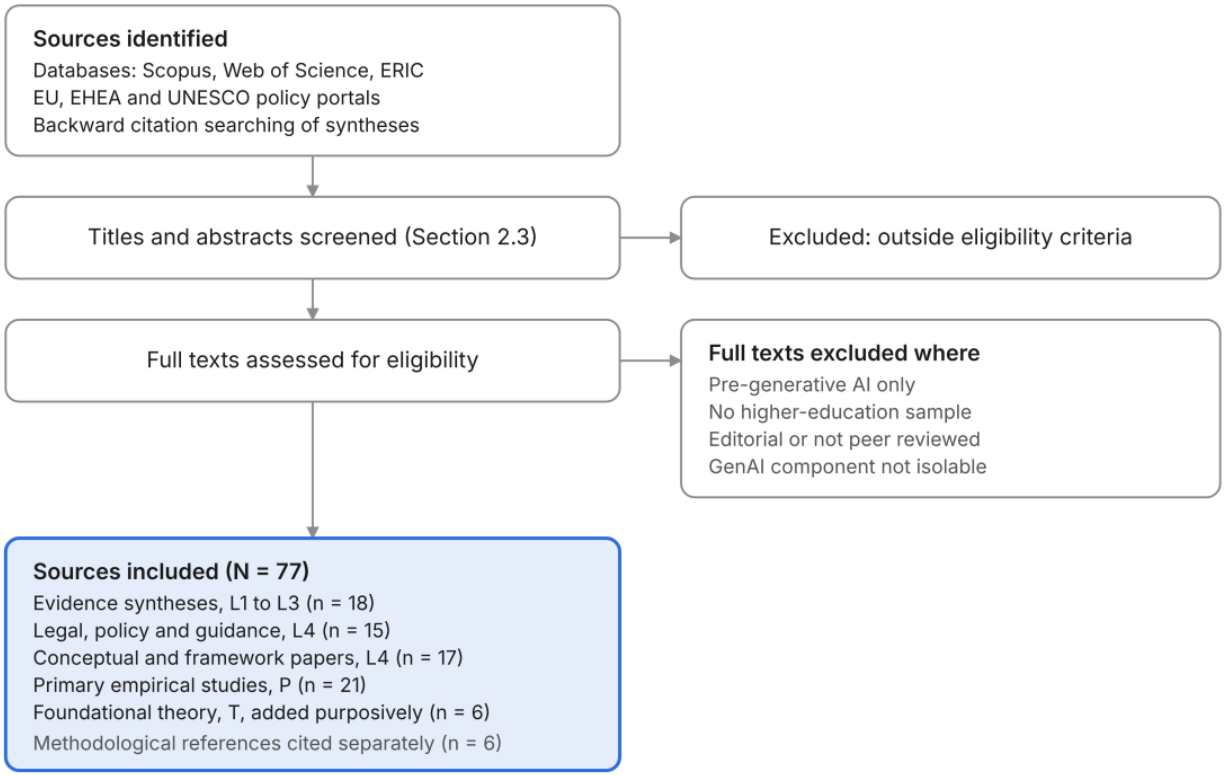


Figure 1. Selection process and composition of the retained corpus ($n = 77$).

2.5. Evidence Classification and the Retained Corpus

Each retained source was classified by type, geographical context and evidence level. Four evidence levels are used: **L1**, meta-analytic evidence; **L2**, systematic reviews of empirical studies; **L3**, mapping and scoping reviews; **L4**, conceptual, legal and policy literature. Primary studies (**P**) and foundational theory (**T**) are reported separately. Throughout Sections 3 to 11, empirical claims are tagged with the level of the evidence that supports them, for example [2] (L1). Claims about learning effectiveness rest on L1 to L3; governance and regulatory analysis rests mainly on L4.

The levels classify sources by design, not by quality, and are not a ranking of evidential weight in the abstract. Three rules govern their use. First, quality is appraised within levels: the ten syntheses that carry the main claims were appraised with AMSTAR-2 or the JBI checklist (Section 2.6), and we report a synthesis rated critically low as indicative even though it sits at L1 or L2. Second, a level determines what a source can support. Claims about learning effects rest on L1 to L3, but a rigorous primary study is given priority over a synthesis when it addresses a question the syntheses cannot answer, such as a natural experiment on assessment validity or an experimental audit of detector bias (Section 8). Third, legal and policy texts are not assessed as empirical evidence. They are used only to support claims about what institutions are required or encouraged to do, are judged by legal status, authority, and currency, and are separated into binding obligations, soft law, and guidance (Section 10). Primary studies were not formally appraised; their designs are recorded in Appendices A and B, and their limitations are stated where they bear on a claim. Table 2 summarizes the retained corpus by source type and evidence level.

Table 2. Retained corpus by source type and evidence level (full list in Appendix A).

Source Type	Evidence Level	Sources (*n*)	European Context (*n*)
Meta-analyses	L1	4	0
Systematic reviews	L2	8	0
Mapping and scoping reviews	L3	6	1
European legal, policy and quality-assurance documents	L4	12	12
International guidance	L4	3	0
Conceptual, critical and framework papers	L4	17	5
Primary empirical studies	P	21	8
Foundational theory	T	6	n/a
Total		77	26

Note. *n* = number of sources; n/a = not applicable (foundational theoretical works are not tied to a geographical context and are therefore not counted in the European-context column).

Six further methodological references [14–19] are cited for method only and are not part of the synthesized corpus.

Two features of the corpus shape all later claims. First, none of the L1 or L2 syntheses is European in scope; all pool global samples dominated by Chinese and US studies [20]. Second, only 8 of 21 primary studies were conducted in European settings. European claims in this review are therefore claims about regulation and policy, supported by a thin layer of context-specific evidence, not claims about "European students" as such.

2.6. Quality Appraisal of Evidence Syntheses

The ten syntheses that carry the review's main claims were appraised with AMSTAR-2 [16] (systematic reviews and meta-analyses) or the JBI checklist for scoping reviews (mapping and scoping reviews). Overall confidence was rated high, moderate, low or critically low using AMSTAR-2's critical domains; publication bias was assessed by funnel-plot asymmetry and Egger's test where reported [19].

Appraisal results are used in two ways. Claims resting on syntheses rated low or critically low are reported as indicative, not established. Where syntheses disagree, the higher rated one is given priority, and the disagreement is reported.

2.7. Synthesis

Synthesis proceeded in three steps. Findings were first coded by evidence level and by the analytical question they address (pedagogical value; risks to learners; institutional implications; sustainability implications; European specificity). They were then mapped onto the four conditions of pedagogical sustainability and the mechanisms involved (scaffolding, cognitive offloading, assessment validity, capability). Finally, findings were compared across streams to identify convergence, contradiction and silence. These comparisons are the basis of the five tensions (Section 12), the framework (Section 13) and the research agenda (Section 15). No separate search for contradictory evidence was run. Instead, sources reporting null, negative or adverse findings were retained whenever they met the criteria and are reported alongside positive findings; Section 12 lists, for each tension, the evidence that qualifies or contradicts it.

2.8. Use of Generative AI Tools

In line with the journal's policy, the use of generative AI in preparing this review is declared here. During revision, Claude Sonnet 5.5 (Anthropic PBC, San Francisco, CA, USA) was used to assist with restructuring and drafting text, with extracting reported methods and statistics from the full texts of the included syntheses for the appraisal (Table 3) and the meta-analytic summary (Section 4.2), and with checking reference formatting and citation order. Grammarly (online version; Grammarly Inc., San Francisco, CA, USA) was used for language editing. All extracted values were checked by the author against the original publications, and all decisions on eligibility, appraisal ratings, interpretation and conclusions were made by the author.

Table 3. Appraisal of the key evidence syntheses (Y = yes; PY = partial yes; N = no).

Synthesis	Type	Studies (*k*)	Protocol Registered	Compre Hensive Search	Risk of Bias Assessed	Appropriate Meta-Analytic Methods	Publication Bias Tested	Overall Confidence
Chen and Cheung [2]	MA	57 (97 effects)	N	PY	PY	PY	Y	Critically low ‡
Fan et al. [21]	MA (three-level)	36 (132 effects)	N	PY	Y	Y	Y	Critically low ‡
Qian [20]	SR	37	N	PY	N	n/a	n/a	Critically low
Baig and Yade-garidehkordi [22]	SR	57	N	PY	N	n/a	n/a	Critically low
Liang et al. [23]	SR	56	N	PY	PY	n/a	n/a	Critically low
Bittle and El-Gayar [24]	SR	41	N	PY	N	n/a	n/a	Critically low
Radovan et al. [25]	SR	23	PY	PY	Y	n/a	n/a	Moderate
Yusuf et al. [1]	Mapping	407	N	PY	n/a	n/a	n/a	Moderate †

Table 3. *Cont.*

Synthesis	Type	Studies (*k*)	Protocol Registered	Compre Hensive Search	Risk of Bias Assessed	Appropriate Meta-Analytic Methods	Publication Bias Tested	Overall Confidence
Cong-Lem and Fan [26]	Scoping	31	N	PY	n/a	n/a	n/a	Low †
Wang et al. [27]	Scoping	123	Y	Y	n/a	n/a	n/a	High †

Note. Y = yes; PY = partial yes; N = no; n/a = not applicable. † For mapping and scoping reviews, overall confidence refers to adherence to JBI and PRISMA-ScR standards; formal risk-of-bias assessment is not expected. ‡ Strict application of AMSTAR-2: two critical domains not met (no registered protocol; no list of excluded studies with reasons) despite otherwise substantive methods. Chen and Cheung [2] searched six databases plus Google Scholar with a reported search string and backward snowballing; controlled bias through inclusion rules (baseline difference below 0.25 SD, randomization where no pretest, at least 15 students per group) rather than a formal tool; used random-effects models but treated 97 effect sizes from 57 studies as independent; and found significant publication bias (Begg's and Egger's tests, trim-and-fill). Fan et al. [21] searched four databases including CNKI with full search strings, used EPHPP-based appraisal in sensitivity analysis, a three-level REML model, and Egger's tests at effect and study level. Qian [20] reports no protocol or quality appraisal and three databases with search terms but not full strings. Baig and Yadegaridehkordi [22] state that methods were defined in advance but report no protocol document or registration; searched ERIC, nine publisher platforms and Google Scholar with a reported string and reference-list searching but excluded gray literature; and applied a four-item checklist on relevance and reporting rather than methodological risk of bias. Liang et al. [23] reported no protocol, searched four databases plus Google Scholar (December 2024) using keyword groups rather than database-specific strings, double-screened with Cohen's $\kappa = 0.72$ and 0.82, and assessed statistical power and reporting rigor rather than risk of bias with a formal tool. Bittle and El-Gayar [24] did not register a protocol and describe no risk-of-bias tool. Radovan et al. [25] used an unregistered a priori protocol, searched four databases with full search strings, and used an MMAT-based appraisal, with single-reviewer extraction. Yusuf et al. [1] report no protocol but followed PRISMA 2020, searched three databases (25 April 2024) with a reported string, double-screened titles, abstracts and full texts (Cohen's $\kappa = 0.81$ and 0.86) and deposited the dataset in OSF; gray literature was not searched. Cong-Lem and Fan [26] report no protocol, searched two databases (to 5 June 2024) with keywords but not full strings, and screened in duplicate in Covidence; their sample spans all educational levels. Wang et al. [27] registered their protocol (INPLASY202650152), searched six databases and screened with three independent reviewers.

3. How the Literature Has Developed

Research on AI in education has moved through three stages. Before 2022, it treated AI mainly as a tool for adaptation, efficiency and automated assessment; a widely cited review found the field dominated by computer science and STEM perspectives, with little pedagogical reflection and little involvement of educators [28] (L2). Early analyses of large language models in education then mapped their opportunities alongside risks such as bias, overreliance and misuse [29] (L4). The release of ChatGPT shifted attention to everyday use: a review of 57 studies from 2023 to 2024 found that adoption, intention-to-use and perception studies based on TAM, UTAUT and UTAUT2 dominated [22] (L2). Current work increasingly asks how GenAI should be integrated, assessed and governed [1,6,7], marking a move from technology adoption to technology stewardship.

The mapping evidence supports the claim that the field is fragmented. Yusuf et al. organized 407 publications into eight discursive themes, led by applications and impacts, ethical risks, user perspectives, adoption and system performance, and identified gaps in experimental and cross-cultural research [1] (L3). Qian similarly organized the empirical literature into separate accounts of pedagogical applications and implementation challenges [20] (L2). Integrative frameworks have appeared since 2025 [6–9], but they connect two or three streams rather than all five that this review examines (Section 13.5).

4. Pedagogical Value and Learning Outcomes

4.1. Scaffolding and Substitution

GenAI is most consistently described as an on-demand scaffold offering explanations, examples and feedback [20,23] (L2). Immediate assistance is not the same as durable learning: a student can submit a better assignment built on a generated explanation while learning little. The decisive variable is how support is positioned. GenAI scaffolds learning when it prompts students to question, verify and build understanding, and substitutes for learning when students delegate the thinking itself [20] (L2). Experimental work comparing direct, reflective and hybrid GenAI feedback reflects this design sensitivity [30] (P). For educators, time saved on drafting materials [31] (L4) is pedagogically valuable only if it is reinvested in dialogue, guidance and assessment design.

4.2. What the Meta-Analytic Evidence Shows, and How Far It Can Be Trusted

Table 4 sets out the meta-analytic estimates on which this review relies, with the information needed to judge them.

Table 4. Meta-analytic estimates of GenAI effects on higher-education learning outcomes.

Synthesis	Studies/Effects	Outcome	Pooled Effect	95% CI	Heterogeneity (I^2)	Publication Bias
Chen and Cheung [2]	57/97	All outcomes	g = 0.804	[0.636, 0.972]	93.1%	Egger $p = 0.002$; trim-and-fill g = 0.321 [0.132, 0.509]
Chen and Cheung [2]	n.r./21	Language skills	g = 2.331	[1.617, 3.044]	96.6%	Egger $p = 0.002$; trim-and-fill g = 1.009
Chen and Cheung [2]	n.r./25	Academic achievement	g = 0.633	[0.418, 0.849]	82.1%	Egger $p = 0.007$; trim-and-fill g = 0.354 [0.111, 0.597]
Chen and Cheung [2]	n.r./25	Affective-motivational	g = 0.617	[0.311, 0.922]	91.3%	Egger $p = 0.96$; none detected
Chen and Cheung [2]	n.r./16	Higher-order thinking	g = 0.580	[0.321, 0.838]	90.2%	Egger $p = 0.86$; none detected
Chen and Cheung [2]	n.r./10	Metacognition	g = 0.078 (n.s.)	[−0.495, 0.652]	94.1%	Egger $p = 0.21$; none detected
Fan et al. [21]	36/132	All outcomes (three-level)	g = 0.499	[0.372, 0.627]	Between-study 23.1%; within-study 0%	Egger $p = 0.13$ (effect level), $p = 0.48$ (study level); no asymmetry
Fan et al. [21]	n.r./43	Higher-order learning (incl. metacognition, self-regulation)	g = 0.504	[0.309, 0.700]	See above	See above
Han et al. [32]	68/337	All outcomes (experimental studies, all levels)	SMD = 0.45	[0.43, 0.47]	95%	Begg $p = 0.157$, Egger $p = 0.224$; fail-safe N = 4882; none detected

Table 4. *Cont.*

Synthesis	Studies/Effects	Outcome	Pooled Effect	95% CI	Heterogeneity (I^2)	Publication Bias
Gu and Yan [33]	19/24	Academic performance (K-12 and higher education)	g = 0.683	[0.17, 1.19]	94.4%	Egger $p = 0.014$ (asymmetry); selection-model adjusted g = 0.681

Note. n.r. = number of studies not reported for this outcome category. Han et al. [32] and Gu and Yan [33] include school as well as university samples.

Three observations follow.

Chen and Cheung's overall estimate is at the upper end of the available syntheses. Independent meta-analyses report average effects between 0.45 and 0.80 [2,21,32,33] (L1). The spread is itself informative: it reflects differences in inclusion criteria, comparators and outcome measures, and it argues against citing any single pooled value as "the" effect of GenAI. In one synthesis, students who received teacher support during their interaction with GenAI gained substantially more (g = 1.426) than those without such support (g = 0.077) [33] (L1), which reinforces the design dependence described in Section 4.1.

The language-skill effect of g = 2.331 should not be read at face value. Effects above 1.0 are very rare in education research. Benchmarks from causal studies place effects above 0.20 on standardized achievement measures among the largest observed [17], and small samples, quasi-experimental designs, and researcher-made measures substantially inflate reported effects, roughly doubling them in some comparisons [18]. Chen and Cheung's own analyses point the same way. Heterogeneity in the language-skill estimate is extreme ($I^2 = 96.6\%$), Egger's test indicates significant small-study asymmetry, and trim-and-fill adjustment reduces the estimate from 2.331 to 1.009; the overall effect falls from 0.804 to 0.321 after the same adjustment [2] (L1). Their moderator analyses also show larger effects for researcher-administered tests, medium-sized samples, 4–12-week interventions and published studies [2]. A further problem is constructing alignment: when language skills are measured by writing tasks completed with GenAI available, the outcome may capture AI-assisted performance rather than learning. We therefore treat 2.331 as an upper bound inflated by small-study and publication effects, not as evidence of a transformative effect.

The metacognition result is a null with uncertainty, not a demonstrated absence of effect. The pooled estimate is g = 0.078, 95% CI [−0.495, 0.652], based on 10 effect sizes with very high heterogeneity ($I^2 = 94.1\%$) [2] (L1). The interval is compatible with no effect and with moderate positive or negative effects, which may differ in direction across instructional designs, and the small number of estimates and the mix of self-report and performance measures limit what can be concluded. Fan et al. report a positive effect for a broader higher-order learning category that includes metacognition and self-regulation (g = 0.504) [21] (L1), which suggests that the null depends partly on how outcomes are grouped. The result still matters: if GenAI performs the monitoring and evaluation that constitute metacognition [34], better task performance can coexist with little growth in self-regulated learning. The appropriate conclusion is that convincing evidence of metacognitive improvement is currently lacking, not that metacognitive benefits are absent or that GenAI weakens independent thinking.

Duration adds a final caution. Effects were stronger in interventions of about 4 to 12 weeks [2] (L1), but no synthesis tests retention after assistance is withdrawn. Research on student use profiles suggests that patterns of use and institutional support are associated with academic outcomes [35] (P), which again points to conditions rather than tools. The relevant question is therefore not whether GenAI improves measured outcomes, but whether the capabilities it develops persist when the tool is removed.

5. Human Agency and Cognitive Offloading

Durable learning depends on capabilities learners can exercise independently: evaluating information, questioning assumptions, constructing arguments, and making decisions [36]. GenAI can act as a cognitive catalyst or enable cognitive offloading, that is, the use of external resources to reduce a task's cognitive demands [37] (T). Offloading is not harmful in itself; it becomes a sustainability problem when it displaces the processes an activity is designed to develop [27] (L3).

The difference lies in how cognitive responsibility is distributed. In a passive sequence (AI → answer → student), the learner accepts or lightly edits an output. In a human-centered sequence (student → AI → critique → verification → reflection → decision), the learner interrogates assumptions, checks claims against evidence and justifies a conclusion. In a three-wave quasi-experimental study, structured contrasts between human and AI performance, supported by guided reflection, were associated with greater career awareness and self-directed learning [38] (P), and making such judgment visible is central to assessment design [39] (L4). GenAI then becomes an object of inquiry rather than an authority, and the learner remains the epistemic decision-maker [40] (L4). This operationalizes the distinction between scaffolding and substitution and grounds the assessment argument in Section 8.

6. From Access to Capability: The GenAI Divide

Digital-divide research has long shown that inequalities persist after physical access becomes universal, shifting to skills and use [41] (T). GenAI adds a capability layer. In a mixed-methods study of 194 university students, Beckman et al. identified novice, cautious and enthusiastic user profiles; students with higher AI literacy reported more constructive uses [42] (P). The study shows that students differ in their capacity to convert access into educational use, not that literacy causes a benefit. A survey of 460 graduate students in China found that perceived usefulness, ease of use and subjective norms directly shaped intention to use GenAI for research, while IT self-efficacy and facilitating conditions acted indirectly [43] (P), which suggests that capability and institutional support shape uptake. Evidence mapping also documents divide phenomena across higher education [44] (L3).

If effective learning depends on critically assessing AI output (Section 5), unequal capability becomes unequal opportunity to benefit. Two further mechanisms deepen the divide: tiered commercial access, in which more capable models sit behind subscriptions (Section 9), and detection tools that misclassify the work of non-native writers (Section 8). In multilingual Europe, both bear directly on equity.

These factors interact rather than simply add up. A student writing in a second language, using assistive technology and unable to afford a paid subscription, may face all three barriers at once, while institutional support, such as embedded literacy teaching and institution-wide licensing, can offset some of them [35,42]. Accessibility is therefore necessary but not sufficient. The social pillar asks for evidence of inclusion, meaning outcomes, access and detector error disaggregated by language, disability and socioeconomic group (Section 13.2), and the maturity model asks institutions to monitor these by student group (Section 14, dimension D5).

7. Evidence from European Higher Education

The European empirical base is small. Of the 21 primary studies retained, 8 were conducted in European settings (Appendix A): a three-institution survey of students and teachers [45] (P), a dual-institution Swedish study of student use [46] (P), an Italian test of an AI Act governance checklist [47] (P), a Norwegian natural experiment on assessment validity [48] (P), a Slovak study of digital competences [49] (P), a multi-country European test of AI-text detection tools [50] (P), a Cypriot–Greek survey of marketing communication

professionals [51] (P) and a qualitative study of European entrepreneurial ventures [52] (P). None of the L1 or L2 syntheses is European in scope, and the overall evidence base is dominated by China and the United States [20] (L2). A scoping review of European university policy responses confirms that institutional evidence is still emerging [53] (L3).

Most European studies are cross-sectional, rely on self-report and use convenience samples, and perceived usefulness is evidence of acceptance, not of durable learning [2,26]. The appropriate conclusion is that European institutions are clearly engaging with GenAI, but the comparative evidence needed to generalize about "European students" or "European universities" does not yet exist. For this reason, the framework in Section 13 is presented as general in architecture and European in its regulatory calibration. Throughout the review, findings from European settings are identified as such; all other empirical claims rest on international evidence interpreted through the European regulatory context and are not presented as established for European students.

8. Academic Integrity, Detection and Assessment Validity

8.1. The Limits of Detection

Institutional responses to GenAI have focused heavily on detecting AI-generated text and reporting misuse [54] (L4). The evidence on detection tools does not support this emphasis. A test of 14 detection tools found them neither accurate nor reliable, with a tendency to classify AI-generated text as human-written [50] (P). Simple adversarial techniques, such as paraphrasing or adding minor errors, substantially reduced detector accuracy further [55] (P). Most consequentially for equity, GPT detectors misclassified more than half of essays written by non-native English writers as AI-generated, while classifying native writers' essays almost perfectly [56] (P).

These findings matter in three ways. For integrity, detection produces both false negatives, so it fails as a deterrent, and false positives, so it can wrongly accuse. For equity, false positives fall disproportionately on multilingual and international students, a large population in European higher education, which turns a technical error rate into a social sustainability problem. For governance, a detection score used to support a misconduct finding influences a student's evaluation. Annex III of the AI Act [57] lists systems intended to monitor and detect prohibited behavior during tests as high-risk, and institutions should at a minimum assess whether their use of detection tools falls into that category. In all cases, detector output should never be the sole basis for an academic decision.

8.2. From Detection to Assessment Validity

The more fundamental question is whether an assessment remains valid when generative tools are available [48] (P). If a system can produce most of a take-home essay, the essay may no longer measure reasoning, synthesis or disciplinary understanding. A systematic review of 41 studies frames GenAI as creating both learning opportunities and integrity risks and identifies digital literacy and institutional responses as priorities [24] (L2); because that review was rated critically low (Table 3), its conclusions are treated here as indicative. The literature converges on process-based and authentic designs that make learning visible: viva-style defences, staged assignments, reflective commentaries, portfolios, authentic tasks and documented AI interactions [6,39,54].

These designs align with sustainable assessment [5] (T): assessment that a machine can complete neither certifies learning nor builds the evaluative judgment on which future learning depends. They do not require excluding GenAI. They assess what humans contribute that cannot be outsourced: judgment, interpretation, verification, ethical reasoning, contextualization and justification of decisions [39]. Because graduates will work alongside

AI systems [51], the goal is to certify both independent capability and the ability to use AI critically and accountably.

Detection, disclosure and assessment redesign are complementary rather than alternatives. Redesign reduces the opportunities for undetectable substitution; clear rules on permitted use and disclosure define what counts as misconduct; and integrity procedures remain necessary for cases that redesign cannot prevent. What the evidence rules out is not monitoring as such but the use of detector scores as decisive evidence. A proportionate approach therefore combines validity-focused design, explicit disclosure requirements, investigation based on several sources of evidence (drafts, process records or an oral discussion of the work) and sanctions proportionate to the breach. The maturity model reflects this combination (Section 14, dimension D3).

9. Governance and the Political Economy of EdTech

9.1. Governance as a System

Recurrent governance concerns include privacy, algorithmic bias, transparency, student well-being and accountability through human oversight [6,58,59] (L4, P). International guidance calls for tools to be validated for ethical and pedagogical appropriateness before adoption [10] (L4). These risks intersect in practice: a single AI-based assessment system may process sensitive data, produce biased evaluations and leave it unclear who is accountable when an automated judgment is wrong [47] (P). Governance therefore has to operate as a system that coordinates academic leadership, teaching, IT, data protection, legal compliance, quality assurance and student support [7,58]. An analysis of documented AI policies across South African public universities shows how such progress can be tracked systematically [60] (P).

9.2. Beyond Vendor Management: The Political Economy of EdTech

Treating vendor dependence as a procurement risk understates the issue. Critical EdTech scholarship shows that the design and deployment of educational technology are shaped by commercial and political interests, not only by pedagogical need [61–63] (L4). Selwyn argues that claims of AI-driven transformation tend to obscure questions about whose interests automation serves and what forms of teaching it displaces [61]. Williamson and colleagues show how datafication and platformization reorganize education around the data and business models of technology providers [62,63]. Komljenovic describes platform providers as rentiers who extract recurring value from control over digital infrastructure and user data [64].

GenAI intensifies these dynamics. The most capable models are developed by a small number of firms, mostly outside Europe, whose pricing, terms of service, data practices and model updates are set unilaterally and can change mid-semester. Three consequences follow for the framework.

- **Equity:** tiered access, in which more capable models sit behind paid subscriptions, creates a pay-to-benefit gradient on top of the capability divide (Section 6). Institution-wide licensing can flatten this gradient, but it also deepens institutional dependence.
- **Integrity:** the market for AI-text detection is itself commercial. Institutions can end up paying one set of vendors to detect the output of another, with tools whose error profile is poorly evidenced (Section 8.1).
- **Governance:** procurement becomes a form of educational policymaking. Contract terms on data use, model training, transparency, portability and exit determine what institutions can govern at all.

The framework therefore treats provider dependence as an institutional and social condition, not a cost line, and the maturity model (Section 14) includes indicators on contract governance, data portability and exit planning.

10. The European Multilevel Policy Context

European governance of GenAI in higher education operates at several levels, not only through EU regulation. Studies of digital education reform and competence development illustrate the practical demands these levels place on institutions [49,65]. Table 5 sets out what each level contributes and what it implies for the framework.

Table 5. Levels of European policy relevant to GenAI in higher education.

Level	Key Instruments	What It Contributes	Implication for the Framework
EU, binding	AI Act (Reg. 2024/1689, amended by Reg. 2026/1744) [57]; GDPR [66]	Binding duties: high-risk status for four listed educational uses only (admission and assignment, evaluation of learning outcomes, assessing the level of education, monitoring tests), with obligations applying from 2 December 2027 at the latest; prohibition of emotion recognition in education and AI literacy duty (Art. 4), both since 2 February 2025; data protection by design and impact assessment (GDPR)	Institutional pillar: compliance, risk classification, human oversight; AI literacy as a legal as well as pedagogical requirement
EU, soft law	Digital Education Action Plan 2021–2027 [67]; Council Recommendation on enabling factors for digital education (2023) [68]; Ethical Guidelines for educators [11]; DigCompEdu [12]; DigComp 2.2 [69]; Coordinated Plan on AI [70]	Shared vision linking technology, competence and inclusion; reference competence frameworks	Social pillar: capability over access; staff-competence indicators
EHEA (Bologna Process)	Rome Communiqué (2020) [71]; Tirana Communiqué (2024) [72]; Standards and Guidelines for Quality Assurance (ESG 2015) [73]	Learning-outcome-based design; student-centered learning and assessment (ESG 1.3); internal quality assurance and ongoing monitoring (ESG 1.1, 1.9); a membership extending beyond the EU	Pedagogical core and institutional pillar: assessment redesign must stay aligned with intended learning outcomes and be auditable under the ESG
National	National AI strategies under the Coordinated Plan; guidance from national quality-assurance agencies; national rules on assessment and examinations	Implementation of EU law; quality-assurance expectations; examination rules; language policy	Calibration: indicators and thresholds set nationally; comparative research (Priority 1)
Institutional	AI policies, assessment regulations, procurement contracts, staff development	Operational governance	Maturity model (Section 14)

The Bologna layer is particularly important and is rarely addressed in the GenAI governance literature reviewed here. Because the ESG require assessment to measure the achievement of intended learning outcomes, assessment validity under GenAI (Section 8.2) is not only a pedagogical concern but a quality-assurance obligation. The EHEA's membership, which extends beyond the EU, also means that the framework cannot rest on

the AI Act alone. The framework is therefore general in architecture, with Table 5 as its European calibration.

Two distinctions follow for the scope of the framework. First, it is intended for higher education in the EHEA, not only the EU. Binding EU law applies directly in EU member states; the GDPR also applies in the other EEA states; and the AI Act can reach providers and deployers outside the EU whose systems produce output used in the Union (Art. 2). Institutions in other EHEA countries are bound by their national law, for which the EU instruments are a reference rather than an obligation. Second, the instruments in Table 5 differ in force. The AI Act and GDPR are binding; the Digital Education Action Plan, the Council Recommendation, and the Ethical Guidelines are policy and guidance; and the ESG are standards that institutions meet through external quality assurance. The framework follows this distinction, treating legal obligations as minimum conditions (Section 13.2) and policy guidance as a reference point.

The AI Act also requires careful qualification, because educational uses of GenAI are not uniformly high-risk. Annex III lists four educational uses as high-risk: determining access, admission or assignment to institutions; evaluating learning outcomes, including when outcomes steer the learning process; assessing the level of education a person will receive or can access; and monitoring and detecting prohibited behavior during tests [57]. Even these can fall outside the high-risk regime where a system performs only a narrow procedural or preparatory task and does not materially influence the outcome of a decision (Art. 6(3)), although the provider must document that assessment and register the system. Most teaching uses, such as tutoring, explanation and feedback on drafts, are not listed, although transparency obligations for certain AI-generated content apply from 2 August 2026 (Art. 50). The prohibition of emotion recognition in educational institutions, except for medical or safety reasons (Art. 5(1)(f)), and the AI literacy obligation (Art. 4) have applied since 2 February 2025. Regulation (EU) 2026/1744, which entered into force in July 2026, postponed the obligations for Annex III high-risk systems from 2 August 2026 to 2 December 2027 at the latest and recast the AI literacy obligation [57]. These dates reflect the legal position at the time of writing.

AI Literacy as a Sustainability Capability

The AI Act's literacy obligation [57], the Ethical Guidelines [11] and the competence frameworks [12,69] all imply that integration requires more than operational skill. An AI-literate student can identify inaccurate output, challenge bias, protect personal data, disclose AI assistance appropriately and recognize when independent judgment is required [74,75] (P). An AI-literate educator can judge when GenAI supports and when it impedes learning, and adapt assessment accordingly [8,76]. DigCompEdu and DigComp 2.2 were not designed for generative tools [12,69], and UNESCO's AI competency framework for teachers [77] offers a more specific reference. Whether literacy initiatives produce lasting gains remains untested (Priority 3).

11. Economic, Environmental and Institutional Conditions

This section summarizes the evidence on the three conditions not covered above; the indicators used to assess them appear in Section 13.2.

Environmental. This is the least developed dimension. A 2026 systematic review found only 23 eligible publications, with electricity, carbon emissions and water use as the main concerns and inconsistent measurement methods [25] (L2). Concerns about the energy cost of training large models predate their classroom use [78] (P). Two distinctions guide the analysis: using AI for sustainability is not the same as making AI sustainable [79], and efficiency in individual tasks can raise total consumption as use expands (a rebound

effect). Environmental claims therefore require a net-impact assessment that the current evidence cannot yet support.

Two kinds of statement should be kept apart. What is established is that GenAI use carries energy, emissions and water costs and that measurement methods are inconsistent [25]. What the framework proposes is an evaluation principle, that resource use should be proportionate to learning value, and applying it faces four practical difficulties. Providers rarely disclose per-query or per-user data, so institutions must rely on estimates. A system boundary must be chosen, for example, whether to include model training, hardware and data-center cooling, or only inference. Shared infrastructure must be allocated between educational and other uses by a stated rule. And comparisons across models and tasks are unreliable, because resource use per query varies with model size and the length of inputs and outputs. The environmental indicators of the SGPF (Section 13.2) are therefore stated as estimates within declared boundaries, and missing provider data are recorded as an information gap rather than estimated away (Section 14.3).

Economic. The price of an individual subscription rarely reflects institution-wide cost, which includes licensing, infrastructure, training, data protection, integration, assessment redesign and, above all, staff time [9,31,80]. Commercial licences concentrate cost in subscriptions and contractual dependence; locally hosted or open-weight models shift it to infrastructure and expertise, with different implications for data protection and environmental accounting. Economic viability depends on whether educational improvements are achieved and sustained relative to total cost of ownership. Cost evidence is thin: the categories above are established, but their size is anticipated rather than documented, because licence prices are observable while no reviewed study has measured staff time, governance, integration, or total cost of ownership. Costs are also likely to vary with institution size, which determines how fixed costs are spread; with the procurement model, from individual subscriptions to institution-wide or consortium licences; with the deployment approach, commercial cloud services or locally hosted open-weight models; and with the use case, since low-stakes tutoring carries fewer governance costs than assessment uses classified as high-risk. Documenting these costs alongside learning outcomes is part of the multi-institution application proposed under Priority 8.

Institutional. Evidence on institution-wide implementation is much thinner than evidence on individual use [7,53] (L4, L3). Research on sustainability in higher education shows that fragmented initiatives become durable only when embedded as strategic capability [9] (L4). Institutional sustainability is therefore treated as an organizational capability and operationalized in the maturity model (Section 14).

12. Five Central Tensions

The five tensions emerged in the third synthesis step (Section 2.7), as points where findings from different streams pull in opposite directions; they were not used to select sources. None can be resolved once and for all, so each must be managed continuously. Table 6 maps each tension to the evidence behind it, the evidence that qualifies or contradicts it, the condition most affected and the management lever.

Table 6. Five central tensions in GenAI-supported higher education.

Tension	Evidence Behind It	Qualifying or Contradictory Evidence	Condition Most Affected	Management Lever
Efficiency vs. cognitive development	Weak evidence of metacognitive benefit; offloading risk [2,27]	Positive effect on a broader higher-order learning category [21]; large gains with teacher support [33]	Pedagogical core	Scaffolded designs; tests without AI

Table 6. *Cont.*

Tension	Evidence Behind It	Qualifying or Contradictory Evidence	Condition Most Affected	Management Lever
Personalization vs. privacy	Data-intensive personalization vs. GDPR minimization and purpose limitation [66]	Students value personalization but raise privacy concerns [74]; locally hosted models can reduce external data flows (Section 11)	Institutional	Data protection impact assessment; minimized data flows
Access vs. capability	Literacy profiles associated with constructive use [42]; tiered commercial access [64]	Institutional support associated with better outcomes [35]; institution-wide licensing can flatten the access gradient (Section 9.2)	Social	Embedded AI literacy; institution-wide licensing
Innovation vs. integrity	Unreliable and biased detectors [50,55,56]; threatened assessment validity [48]	Validity can be protected through process-based and authentic designs that permit AI use [6,39]	Pedagogical core and institutional	Process-based assessment; disclosure rules; no sole reliance on detectors
Scalability vs. environmental responsibility	Thin, inconsistent measurement; rebound effects [25]	Per-task efficiency gains are possible, but net effects are unmeasured; using AI for sustainability differs from making AI sustainable [79]	Environmental	Proportionate use; disclosure requirements in procurement

13. The Sustainable GenAI Pedagogy Framework (SGPF)

13.1. Architecture

The SGPF translates the definition in Section 1.1 into an evaluation sequence (Figure 2). GenAI capabilities enter education through pedagogical integration, defined by learning objectives, instructional design and rules for appropriate use. The integration must then pass a gate: is there evidence of durable learning value, meaning retention, transfer or independent capability, not only task performance? If not, the practice is redesigned or stopped, whatever its cost, accessibility or carbon profile. If it passes, it is weighed against four condition pillars. Monitoring evidence returns to design in a continuous-improvement loop, and five principles constrain every stage.

The framework is general in architecture: the gate, pillars and principles apply to any higher education system. It is European in calibration: the institutional pillar's indicators are anchored to the AI Act, GDPR and ESG 2015 (Table 5), and its social indicators to multilingual and cross-border student populations.

13.2. Condition Pillars and Indicators

Table 7 states the pillars, their guiding questions and indicators, how each is measured, and the minimum condition each must meet, with the evidence that motivates it.

Weighing the pillars. The pillars are neither equally weighted nor fully compensable. Each has a minimum condition that must be met before any trade-off is considered, and the institutional minimum is a legal floor rather than a matter of judgment. Above these floors, the relative weight of the pillars is context-dependent and is set by the institution, with national quality-assurance expectations as the reference (Table 5). The current evidence cannot yield numerical thresholds, such as an acceptable cost per learner or emissions estimate, so institutions and national authorities set them through calibration. Trade-offs above the floors are resolved by documented deliberation: the panel that applies the maturity model (Section 14.2) records which pillar was favored, why, and what mitigation applies

to the others. Consider a tool that improves accessibility for disabled students but raises cost and computational demand. If it passes the learning-value gate and meets all four minimum conditions, the social gain may justify the additional cost and energy, provided the smallest adequate model is used and the decision is reviewed when monitoring data arrive. If it meets the social floor only by breaching another, for example, by processing student data without an impact assessment, it is not adopted until that breach is remedied.

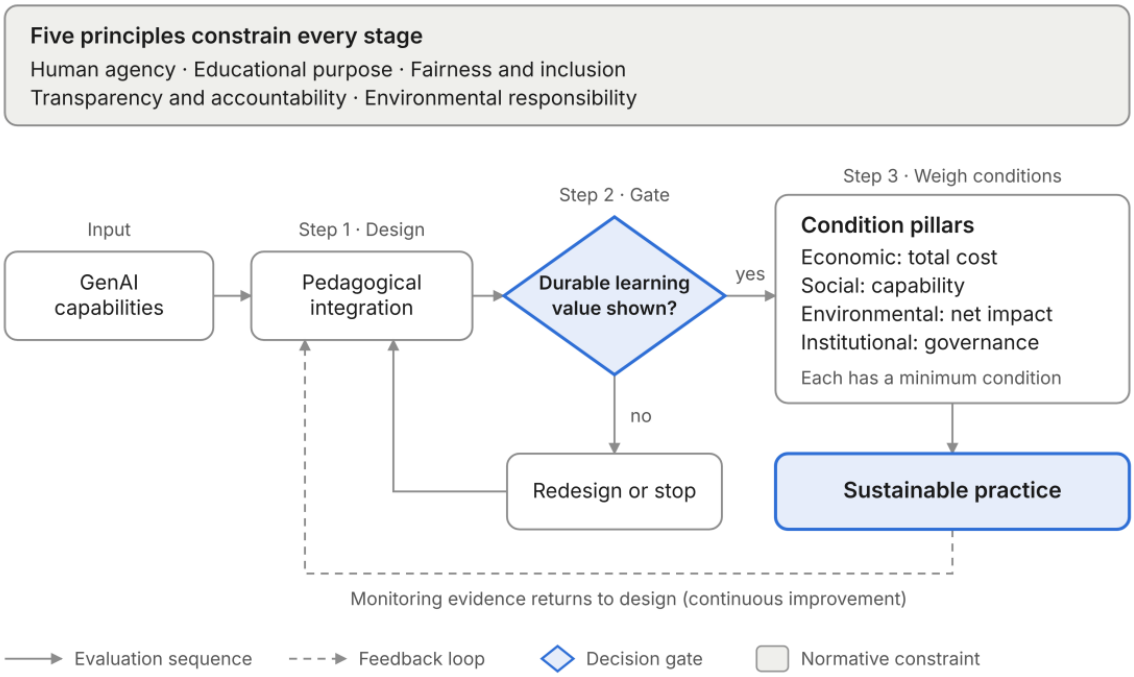


Figure 2. Sustainable GenAI Pedagogy Framework (SGPF). Durable learning value acts as a gate; the four condition pillars, each with a minimum condition, weigh only practices that pass it, and five principles constrain every stage. Solid arrows show the evaluation sequence and the dashed arrow the feedback loop.

Table 7. SGPF condition pillars: guiding questions, indicators, measurement and minimum conditions.

Pillar	Guiding Question	Indicators	How Measured	Minimum Condition	Evidence Base
Economic	Does the learning value justify total cost?	Total cost of ownership; cost per learner; staff time; infrastructure; exit and switching costs	Cost accounting per learner per term over the contract period, including staff time	Total cost known and funded for the full contract period, including exit	Section 11 [9,31,80]
Social	Who gains, and who is left behind?	Measured AI literacy by group; disability and language accessibility; access tier; false-positive rates of any detection tool by student group	Performance-based literacy test; outcomes, access and detector error disaggregated by language, disability and socioeconomic group	No student group measurably worse off than without the practice; accessible to disabled students	Sections 6 and 8.1 [42–44,56]

Table 7. *Cont.*

Pillar	Guiding Question	Indicators	How Measured	Minimum Condition	Evidence Base
Environmental	Are resource demands proportionate to learning value?	Estimated energy, emissions and water per use; model size relative to task; intensity of use; provider disclosure	Provider-disclosed or estimated energy per use within a declared boundary	Smallest adequate model used; data gaps recorded and raised in procurement	Section 11 [25,78,79,81]
Institutional	Can the practice be governed and sustained?	AI Act risk classification; data protection impact assessment; ESG-aligned assessment review; staff competence; contract terms on data, portability and exit	Documentary evidence: risk classification, DPIA, assessment review, contract terms	Legal compliance (AI Act, GDPR) and ESG-aligned assessment; not tradeable	Sections 9 and 10 [47,53,57,73]

13.3. Principles

Five principles supply the normative criteria for weighing trade-offs between pillars. *Human agency:* learners and educators retain responsibility for evaluation and decisions [27,38,40]. *Educational purpose:* use responds to a defined learning need, not to novelty or efficiency alone [2,31]. *Fairness and inclusion:* implementation addresses gaps in access, proficiency, language and disability [42–44]. *Transparency and accountability:* expectations for use, disclosure, data processing and oversight are explicit [57,58]. *Environmental responsibility:* use is proportionate to purpose [25,79].

13.4. Theoretical Grounding and Propositions

The framework draws on three traditions. The triple bottom line holds that economic, social and environmental performance must be assessed jointly [82]; the SGPF adds an institutional dimension reflecting AI's governance demands. The capability approach distinguishes access to resources from the real freedom to convert them into valued functionings [83], which underpins the move from access to capability (Section 6) and has been applied to learner agency in GenAI-supported education [40]. Sustainable assessment [5] grounds the durability criterion at the gate. None of these traditions is new, and the SGPF does not claim new components. Its contribution lies in how they are combined: learning value becomes a non-compensable gate rather than one consideration among several; each pillar carries a minimum condition, so that the triple bottom line's dimensions cannot simply be traded against one another; and the whole is tied to binding European obligations and to an operational maturity model (Section 13.5).

Three propositions follow for empirical testing. **P1:** the sustainability of GenAI-supported practice depends more on pedagogical design than on tool capability. **P2:** the distribution of benefit is shaped by learner capability, institutional support and access tier. **P3:** long-term viability depends on institutional and environmental conditions that immediate performance metrics do not capture. These are hypotheses derived from cross-stream synthesis, not established causal relationships.

13.5. Comparison with Existing Frameworks

Table 8 compares the SGPF with eight frameworks and models. Ratings reflect each document's stated scope.

Table 8. The SGPF compared with existing guidance, competence frameworks, maturity models and integrative frameworks (Y = central; P = partial; N = not addressed).

Framework	Unit of Analysis	Learning Value as Criterion	Assessment Validity	Equity and Capability	Environmental	Economic (TCO)	Regulatory Anchoring	Progression with Indicators
UNESCO guidance on GenAI (2023) [10]	System and institution	P	P	Y	N	N	P (international)	N
EU Ethical Guidelines for educators (2022) [11]	Educator	P	P	Y	P	N	P (pre-AI Act)	N
DigCompEdu (2017) [12]	Educator competence	P	Y	P	N	N	N	Y (individual proficiency)
Jisc AI maturity model (2024) [13]	Institution	N	P	P	N	N	N	Y (five stages)
Sejdiu et al. (2026) [6]	Pedagogy, assessment, governance	P	Y	P	N	N	P	N
Kuang and Gao (2026) [7]	Institution, teaching, student	P	P	P	N	N	N	N
Torres-Rivera et al. (2025) [8]	Curriculum (STEAM)	P	N	P	P	N	N	N
Sakka et al. (2026) [9]	Institution (sustainability, not AI-specific)	N	N	P	Y	P	N	P
SGPF (this review)	Practice and institution	Y (non-compensable gate)	Y	Y	Y	Y	Y (AI Act, GDPR, ESG)	Y (Section 14)

The comparison shows where the SGPF adds to existing work. No compared framework combines economic and environmental accounting with pedagogy and governance; none treats learning value as a necessary condition rather than one consideration among several; and only the Jisc model and DigCompEdu specify progression, at institutional and individual levels respectively. The SGPF does not replace these instruments but is designed to sit above them: DigCompEdu and UNESCO's teacher framework [77] can supply staff-competence indicators, the Ethical Guidelines the questions behind the transparency principle, and the Jisc stages a cross-check for the maturity levels.

14. Operationalizing the Institutional Pillar: A GenAI Maturity Model

14.1. Levels and Dimensions

The maturity model operationalizes the institutional pillar and makes the framework usable for self-assessment and benchmarking. It has five levels (Figure 3) assessed across six dimensions (Table 9). Its stages were cross-checked against the Jisc AI maturity model, which uses five stages from approaching and understanding to optimized or transformed [13]. The

SGPF model differs in three ways: it adds environmental, economic and equity dimensions; it ties governance indicators to the AI Act, GDPR and ESG; and it specifies a scoring rule.

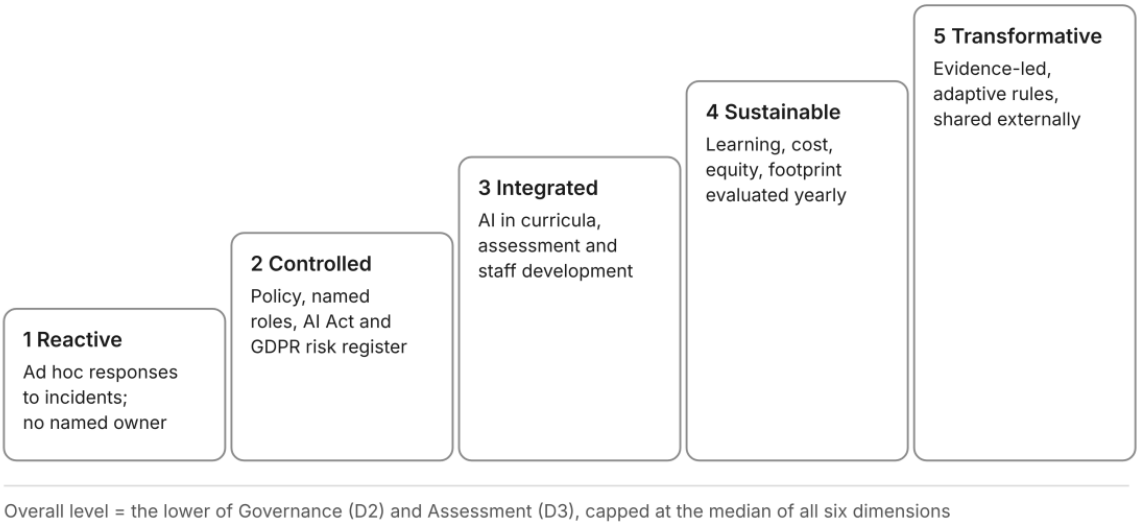


Figure 3. Institutional GenAI Maturity Model: five levels assessed across six dimensions (Table 9). The overall level is the lower of the two critical dimensions (D2, D3), capped at the median of all six.

Table 9. Indicators by dimension and level. A level is awarded only when its descriptor and all lower descriptors are evidenced by a document or dataset.

Dimension	1 Reactive	2 Controlled	3 Integrated	4 Sustainable	5 Transformative
D1 Strategy and evaluation	No strategy; responses to incidents	Approved AI policy with named owner and review date	AI objectives in institutional strategy; annual policy review	Annual evaluation of learning outcomes and costs informs decisions	Evaluation published; findings shared with sector
D2 Governance and compliance (critical)	No inventory of AI uses	Inventory of uses; AI Act risk classification; DPIAs for high-risk uses	Human-oversight procedures for all high-risk uses; incident log	Periodic audit of high-risk uses; contract terms on data, portability and exit	Independent audit; student participation in governance
D3 Pedagogy and assessment (critical)	Ban or silence; reliance on detectors	Programme-level rules on permitted AI use; detectors never sole evidence	Assessment redesigned and mapped to intended learning outcomes (ESG 1.3)	Validity of redesigned assessment monitored, including tests without AI	Assessment innovation evaluated and shared
D4 Capability and AI literacy	Optional ad hoc workshops	Baseline literacy offer for staff and students (AI Act Art. 4)	Discipline-specific literacy in curricula and staff development	Literacy measured by performance, not attendance	Literacy outcomes benchmarked externally
D5 Equity and access	Unequal, unmonitored access	Institution-wide access to at least one vetted tool	Accessibility and language support reviewed	Outcomes and detector false-positive rates monitored by student group	Gaps closed and reported
D6 Resources, environment and providers	Spending unknown	Total cost recorded; single-provider risk identified	Total cost of ownership per learner; exit plan	Environmental disclosure required in procurement; proportionality reviewed	Net-impact accounting published

14.2. Scoring Rules

1. **Evidence:** each dimension is rated 1 to 5 by a panel including quality assurance, academic staff, the data protection officer and a student representative. Each rating must cite a document, contract or dataset, and inter-rater agreement is reported.
2. **Cumulative descriptors:** a level is awarded only if all lower-level descriptors are also met.
3. **Overall level:** the lower of D2 and D3, capped at the median of all six dimensions, rounded down. Governance and assessment are critical because they cover the uses the AI Act classifies as high-risk and the ESG requires to be valid.
4. **Profile, not only a number:** results are reported as a six-dimension profile, so that a single score does not hide a weak dimension.

These thresholds are a theoretical starting point. Their validity, including whether progression predicts better learning and equity outcomes, must be tested empirically (Priority 8).

Three design choices also require justification. The five levels follow the convention of capability maturity models and allow a direct cross-walk to the five Jisc stages [13]; fewer levels would merge distinct stages, such as the difference between having rules (level 2) and having redesigned practice (level 3), while more would require distinctions the current evidence cannot support. Progression is ordered but not strictly linear: descriptors are cumulative within a dimension, but dimensions can advance at different speeds, so results are reported as a profile, and an institution can fall back, for example, after a change in provider or a new legal obligation. Movement between levels is demonstrated only by new documentary evidence at reassessment, normally annually, against the descriptors in Table 9, each of which names an observable practice, such as an inventory of uses, a DPIA, a performance-based literacy measure or published net-impact accounting.

14.3. Illustrative Application

The following case is constructed for illustration and does not report empirical data.

Context. A mid-sized private university in Southern Europe plans to deploy a GenAI feedback assistant on draft assignments in a second-year undergraduate marketing module with about 180 students, many of them international students writing in a second language.

Step 1: The gate. The module team states the intended learning outcome (applying segmentation frameworks to an unfamiliar market) and designs the evidence the gate requires: a comparison between a cohort with the assistant and a prior cohort, plus an unaided in-class case analysis in week 10 to test whether gains persist without the tool. Until that evidence exists, the deployment is a pilot, not sustainable practice.

Step 2: The pillars. Economic: the licence is a small share of cost; the larger cost is staff time to redesign the assignment and review feedback quality, and the total cost for the pilot year, including the cost of switching provider, is estimated and funded. Social: AI literacy is uneven, so the institution embeds a two-hour verification workshop in week 2 and decides not to use AI-text detectors, given their false-positive rates for non-native writers [56]. Environmental: the smallest model tier adequate for feedback on drafts is selected, and the provider discloses no energy data, which is recorded as an information gap and raised at contract renewal. Institutional: feedback on drafts does not itself evaluate learning outcomes, so it is classified outside Annex III, but a data protection impact assessment is completed because student texts are processed. The minimum conditions in Table 7 are therefore met, except the social condition, which depends on outcome data and is built into the pilot's evaluation by comparing results across language groups; the remaining trade-off, the staff time required, is judged acceptable for a pilot and recorded for review.

Step 3: Maturity profile. Table 10 sets out the illustrative rating for each dimension and the evidence cited for it.

Table 10. Illustrative maturity profile for the constructed case.

Dimension	Illustrative Rating	Evidence Cited
D1 Strategy and evaluation	3	AI objectives in the strategic plan; annual policy review
D2 Governance and compliance	2	Inventory, AI Act risk classification and DPIAs exist; no human-oversight procedure yet
D3 Pedagogy and assessment	3	Module assessment mapped to learning outcomes
D4 Capability and AI literacy	2	Baseline workshops; no performance measure
D5 Equity and access	3	Institution-wide licence; accessibility review
D6 Resources, environment and providers	2	Total cost recorded; no exit plan or environmental data

The median is 2.5, rounded down to 2, and the lower of D2 and D3 is 2, so the overall level is 2 (Controlled). The profile shows why: assessment and strategy are ahead of governance, literacy measurement and provider management. The priority actions are a human-oversight procedure (D2), a performance-based literacy measure (D4) and contract terms on portability and environmental disclosure (D6). The case shows how the gate and the maturity model work together: the module-level practice may pass the gate, while the institution's capacity to sustain it remains at level 2.

15. Research Agenda

Three commitments apply to every priority: measure learning directly, not only perceptions; report the GenAI tools, versions and conditions of use, given how quickly systems change; and, where feasible, preregister designs and share instruments to allow replication across European contexts. Table 11 sets out the research priorities, the gap each addresses and suggested designs.

For Priority 6, this review proposes a conceptual heuristic, not a quantitative measure, for comparing deployments:

$$\textit{Pedagogical sustainability efficiency} = \frac{\textit{Demonstrable durable learning value}}{\textit{Computational resource cost}}$$

The ratio is meaningful only under three restrictions. First, it compares alternatives for the same learning outcome, student population, and assessment unit, such as two configurations of a feedback assistant in one module; it cannot rank deployments across tasks or disciplines. Second, the numerator is the difference on a validated, unaided measure of the intended outcome, preferably delayed, between the configuration and a comparison condition; perceptions and scores obtained with AI available do not count. Third, the denominator is an estimate of energy or emissions per student for the module within a declared boundary, for example, inference only, with training and shared infrastructure excluded or allocated by a stated rule. The two terms remain in different units and are never combined with other pillars into a single score. Used in this way, the ratio supports an ordinal judgment: if a smaller model achieves the same unaided learning gain as a larger one, the smaller model is preferred; if the larger model achieves a greater gain, the choice returns to the deliberation described in Section 13.2. Until providers disclose comparable data (Section 11), the denominator will usually be an order-of-magnitude estimate.

Table 11. Research priorities, the gap each addresses and suggested designs.

Priority	Gap Addressed	Suggested Design
1. Comparative European research	Eight European primary studies; no cross-national comparisons (Section 7)	Coordinated cross-national panels across Nordic, Western, Southern, Central and Eastern systems, with common measures
2. Longitudinal learning	No synthesis tests retention after assistance is withdrawn (Section 4.2)	Multi-semester cohorts; repeated assessments with and without AI
3. AI literacy measurement	Literacy measured by self-report (Sections 6 and 10)	Performance-based instruments validated across languages and student groups
4. Assessment redesign	Validity of AI-integrated designs untested (Section 8.2)	Quasi-experiments comparing product-only and process-documented designs
5. Detection and fairness	Detector error and bias in multilingual settings (Section 8.1)	Audits of false-positive rates by first language, discipline and level
6. Environmental accounting	23 studies, inconsistent metrics (Section 11)	Standard reporting of energy, emissions and water per use, linked to learning value
7. Multilingual and accessible GenAI	Equivalence of support across languages and disabilities unknown (Section 6)	Standardized tasks and language-specific benchmarks; participatory designs with disabled students
8. Validation of the SGPF and maturity model	Framework and thresholds are conceptual (Sections 13 and 14)	Expert Delphi on indicators; multi-institution application; tests of whether maturity predicts outcomes

16. Implications for Policy and Practice

For university leaders, quality-assurance agencies and the wider European education market, the analysis implies six priorities. Their evidential basis differs, and each is labeled accordingly: E, supported by empirical evidence reviewed here, although no reviewed study tests the recommendation itself as an intervention; L, required by law; N, a theoretically or normatively grounded recommendation.

1. **Evaluate GenAI by durable learning, not adoption.** Apply the gate before scaling any use (N, grounded in the variability of effects reported in Section 4).
2. **Embed governance in pedagogy.** Coordinate AI policy with curriculum, assessment, data protection and quality assurance, and classify uses under the AI Act (L for classification and data protection; N for coordination).
3. **Treat AI literacy as an equity measure.** Teach verification, bias recognition, privacy and disclosure, and measure competence rather than attendance (L, AI Act Art. 4; E for the association between literacy and constructive use, which is correlational).
4. **Redesign assessment around validity.** Use process-based and authentic designs aligned with intended learning outcomes, and never use detector output as sole evidence of misconduct (E for detector error and bias; N for redesign).
5. **Govern providers, not only products.** Negotiate data use, portability, exit and environmental disclosure, recognizing procurement as educational policy (N).
6. **Prepare graduates for AI-augmented professional competence.** This means using AI while retaining independent judgment and accountability [51,52], without letting short-term employer demand displace education's broader purposes (N).

17. Limitations and Conclusions

17.1. Limitations

This is an integrative, not a systematic, review of primary studies. Sources were selected purposively against stated eligibility criteria, and records were not logged at each stage of

the search and screening, so the selection cannot be reproduced exactly, and stage-by-stage record counts are not reported. Table 1 provides a structured search string so that the selection can be replicated and updated. The purposive inclusion of foundational theory and policy texts is a further source of selection bias. Restricting the review to English may underrepresent research published in other European languages, which is significant given the review's European focus. Because the review relies mainly on secondary syntheses, its conclusions inherit their limitations. The syntheses likely overlap in the primary studies they include, so their agreement reflects shared sources rather than independent confirmation; they vary in quality, and most were rated critically low (Table 3); and none is European in scope. Selective inclusion means that unpublished, non-English and null findings may be under-represented, which would bias the synthesis towards positive effects. The European conclusions rest on eight primary studies and on regulatory analysis and should be read as provisional. The appraisal relies on published synthesis reports and was carried out by a single reviewer, increasing the risk of appraisal error. Both the SGPF and the maturity model are conceptual; the illustrative application demonstrates use, not validity. The field and its regulation are changing rapidly.

17.2. Conclusions

The evidence indicates that GenAI can support achievement, motivation and higher-order thinking, with moderate average effects that vary widely by design and context. The very large language-skill effect is not robust, convincing evidence that GenAI improves metacognition is lacking, and durability is untested. Educational value must therefore be distinguished from task efficiency and perceived usefulness.

The review contributes a single definition of pedagogical sustainability; a framework in which durable learning value is a non-compensable gate weighed against economic, social, environmental and institutional conditions; a comparison showing what the framework adds to existing guidance and models; and an operational maturity model with indicators, scoring rules and an illustrative application. It also shows that equity, integrity, and governance arguments depend on two often-overlooked issues: the unreliability and bias of AI-text detection, and the political economy of a provider market concentrated in a few firms.

Sustainable integration requires evidence-informed stewardship rather than adoption or compliance alone. The challenge for European higher education is not simply to adopt GenAI, but to create the conditions under which its learning benefits can be demonstrated, shared fairly and sustained.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article; the appraisal forms are available from the author on request.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AI	Artificial intelligence
AMSTAR-2	A MeaSurement Tool to Assess systematic Reviews, version 2
CI	Confidence interval
DPIA	Data protection impact assessment

EEA	European Economic Area
EHEA	European Higher Education Area
ESD	Education for sustainable development
ESG	Standards and Guidelines for Quality Assurance in the European Higher Education Area
EU	European Union
GDPR	General Data Protection Regulation
GenAI	Generative artificial intelligence
JBI	Joanna Briggs Institute
L1–L4	Evidence levels 1 to 4 (Section 2.5)
P	Primary empirical study
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SGPF	Sustainable GenAI Pedagogy Framework
SMD	Standardized mean difference
T	Foundational theory
TAM	Technology acceptance model
TCO	Total cost of ownership
UTAUT	Unified theory of acceptance and use of technology

Appendix A. Retained Sources by Type, Context and Evidence Level

Table A1 lists the retained sources by type, context and evidence level. Context is the setting of the data or of the regulation, not author affiliation. Methodological references [14–19] are excluded.

Table A1. Retained sources by type, context and evidence level.

Ref.	Source	Type	Level	Context
[1]	Yusuf et al. 2024	Mapping review (407 publications)	L3	Global
[2]	Chen and Cheung 2025	Meta-analysis (57 studies)	L1	Global
[3]	UNESCO 2020	International guidance	L4	International
[4]	Bianchi et al. 2022	EU framework (GreenComp)	L4	EU
[5]	Boud 2000	Theory	T	n/a
[6]	Sejdiu et al. 2026	Integrative framework	L4	n/a (conceptual)
[7]	Kuang and Gao 2026	Integrative framework	L4	n/a (conceptual)
[8]	Torres-Rivera et al. 2025	Conceptual model	L4	n/a (conceptual)
[9]	Sakka et al. 2026	Conceptual	L4	n/a (conceptual)
[10]	Miao and Holmes 2023	International guidance (UNESCO)	L4	International
[11]	European Commission 2022	EU guidance	L4	EU
[12]	Redecker 2017	EU framework (DigCompEdu)	L4	EU
[13]	Jisc 2024	Sector maturity model	L4	United Kingdom
[20]	Qian 2025	Systematic review (37 studies)	L2	Global
[21]	Fan et al. 2026	Three-level meta-analysis	L1	Global
[22]	Baig and Yadegaridehkordi 2024	Systematic review (57 studies)	L2	Global
[23]	Liang Z. et al. 2026	Systematic review	L2	Global
[24]	Bittle and El-Gayar 2025	Systematic review (41 studies)	L2	Global
[25]	Radovan et al. 2026	Systematic review (23 studies)	L2	Global
[26]	Cong-Lem and Fan 2027	Scoping review	L3	Global
[27]	Wang et al. 2026	Scoping review	L3	Global

Table A1. *Cont.*

Ref.	Source	Type	Level	Context
[28]	Zawacki-Richter et al. 2019	Systematic review	L2	Global
[29]	Kasneci et al. 2023	Conceptual	L4	Global
[30]	Ateş 2026	Multisite experiment	P	Four universities (country not stated)
[31]	Giannakos et al. 2025	Conceptual, expert	L4	Europe
[32]	Han et al. 2025	Meta-analysis	L1	Global
[33]	Gu and Yan 2025	Meta-analysis (19 studies)	L1	Global
[34]	Flavell 1979	Theory	T	n/a
[35]	Zhang et al. 2026	Survey ($n = 2240$)	P	China
[36]	Felix et al. 2025	Systematic review	L2	Global
[37]	Risko and Gilbert 2016	Theory	T	n/a
[38]	Jou and Hao 2026	Quasi-experimental mixed methods ($n = 176$)	P	Taiwan (inferred from author affiliation and funding)
[39]	Sharma 2026	Conceptual	L4	Global
[40]	Saito 2026	Conceptual	L4	Global
[41]	van Dijk 2005	Theory	T	n/a
[42]	Beckman et al. 2025	Mixed methods ($n = 194$)	P	Australia
[43]	Yu et al. 2025	Survey ($n = 460$; PLS-SEM, fsQCA)	P	China
[44]	Gide and Al-Sabbagh 2026	Evidence map	L3	Global
[45]	Žáková et al. 2025	Survey, three institutions	P	Europe
[46]	Nygren et al. 2026	Dual-institution study	P	Sweden
[47]	Manganello et al. 2025	Applied governance study	P	Italy
[48]	Bratli et al. 2026	Natural experiment	P	Norway
[49]	Marasová et al. 2024	Survey	P	Slovakia
[50]	Weber-Wulff et al. 2023	Tool evaluation	P	Europe
[51]	Komodromos et al. 2024	Survey	P	Cyprus, Greece
[52]	Komodromos 2026	Qualitative chapter	P	Europe
[53]	Scierri et al. 2026	Scoping review	L3	Europe
[54]	Peterson 2025	Theoretical opinion article	L4	n/a (conceptual)
[55]	Perkins et al. 2024	Experimental evaluation	P	International
[56]	Liang W. et al. 2023	Experimental audit	P	USA
[57]	Regulation (EU) 2024/1689	EU law (AI Act)	L4	EU
[58]	Crompton et al. 2026	Delphi	P	Global
[59]	Kim et al. 2025	Legal-conceptual	L4	EU
[60]	Jacobs-Basadien and Pather 2026	Document analysis	P	South Africa
[61]	Selwyn 2019	Critical scholarship	L4	Global
[62]	Williamson and Eynon 2020	Critical scholarship	L4	Global
[63]	Williamson et al. 2023	Critical scholarship	L4	Global
[64]	Komljenovic 2021	Critical scholarship	L4	Global
[65]	Vakulyk et al. 2023	Policy analysis	L4	EU, UK

Table A1. *Cont.*

Ref.	Source	Type	Level	Context
[66]	Regulation (EU) 2016/679	EU law (GDPR)	L4	EU
[67]	European Commission 2020	EU policy (DEAP)	L4	EU
[68]	Council of the EU 2023	EU recommendation	L4	EU
[69]	Vuorikari et al. 2022	EU framework (DigComp 2.2)	L4	EU
[70]	European Commission 2021	EU policy	L4	EU
[71]	EHEA 2020	Ministerial communiqué	L4	EHEA
[72]	EHEA 2024	Ministerial communiqué	L4	EHEA
[73]	ENQA et al. 2015	Quality-assurance standards (ESG)	L4	EHEA
[74]	Shen et al. 2025	Qualitative	P	China
[75]	Lin et al. 2025	Survey	P	China
[76]	Gibreel et al. 2026	Survey	P	Country not stated (437 undergraduates)
[77]	Miao and Cukurova 2024	International framework (UNESCO)	L4	International
[78]	Strubell et al. 2019	Technical study	P	USA
[79]	Cañavate et al. 2025	Conceptual, case	L4	Spain
[80]	Peng and Li 2025	Systematic review	L2	Global
[81]	Esho et al. 2026	Review	L3	Global
[82]	Elkington 1997	Theory	T	n/a
[83]	Sen 1999	Theory	T	n/a

Note. n = sample size; n/a = not applicable (foundational theoretical works, and conceptual sources that do not draw on data from, or regulate, a particular setting).

Appendix B. Evidence Matrix for the Empirical Sources

The matrix in Table A2 covers the 18 evidence syntheses and 21 primary studies. Findings are those used in this review; synthesis limitations draw on the appraisal in Table 3, and primary-study limitations follow from their design.

Table A2. Evidence matrix for the empirical sources.

Ref.	Design	Context	Finding Used in This Review	Main Limitation	Relevance to the SGPF
[1]	Mapping review (407 publications)	Global	Eight discursive themes; few experimental studies; limited attention to culture	No appraisal of included studies; gray literature not searched	Fragmentation claim (Sections 1 and 3)
[2]	Meta-analysis (57 studies, 97 effects)	Global	Overall g = 0.804; language skills g = 2.331; metacognition g = 0.078 (n.s.); stronger effects at 4–12 weeks	Critically low; publication bias; dependent effects treated as independent	Learning-value gate (Section 4.2)
[20]	Systematic review (37 studies)	Global	Scaffolding vs. substitution; evidence dominated by China and the USA	Critically low; no quality appraisal	Pedagogical core (Sections 2.5 and 4.1)

Table A2. *Cont.*

Ref.	Design	Context	Finding Used in This Review	Main Limitation	Relevance to the SGPF
[21]	Three-level meta-analysis (36 studies)	Global	g = 0.499; higher-order learning g = 0.504	Critically low under AMSTAR-2 critical domains	Learning-value gate (Section 4.2)
[22]	Systematic review (57 studies)	Global	Adoption and intention studies (TAM, UTAUT) dominate	Critically low; reporting checklist only	Field development (Section 3)
[23]	Systematic review (56 studies)	Global	GenAI as an on-demand scaffold	Critically low; no formal risk-of-bias tool	Pedagogical core (Section 4.1)
[24]	Systematic review (41 studies)	Global	Learning opportunities alongside integrity risks	Critically low; no risk-of-bias tool	Assessment validity (Section 8.2)
[25]	Systematic review (23 publications)	Global	Electricity, emissions and water as main concerns; inconsistent measurement	Moderate; single-reviewer extraction	Environmental pillar (Section 11)
[28]	Systematic review	Global	Pre-2022 AI-in-education research dominated by computer science, little pedagogy	Pre-generative AI	Field development (Section 3)
[32]	Meta-analysis (68 studies)	Global	SMD = 0.45 across experimental studies	Not appraised; includes school samples	Triangulation of effects (Section 4.2)
[33]	Meta-analysis (19 studies)	Global	g = 0.683; teacher support g = 1.426 vs. 0.077 without	Not appraised; includes school samples; small-study asymmetry	Design dependence (Section 4.2)
[36]	Systematic review	Global	Critical thinking as an independent capability	Not specific to GenAI	Human agency (Section 5)
[80]	Systematic review	Global	Cost components of AI-supported personalized learning	Not a cost study	Economic pillar (Section 11)
[26]	Scoping review (31 studies)	Global	Perceived usefulness is not evidence of durable learning	Low; all educational levels	European evidence (Section 7)
[27]	Scoping review (123 studies)	Global	Offloading becomes a problem when it displaces targeted processes	High (JBI)	Human agency principle (Section 5)
[44]	Evidence map	Global	Divide phenomena across higher education	Not appraised	Social pillar (Section 6)
[53]	Scoping review	Europe	European institutional policy evidence is still emerging	Not appraised; published research only	Institutional pillar (Sections 7 and 11)
[81]	Review	Global	Environmental impact of GenAI	Not appraised	Environmental indicators (Table 7)
[30]	Multisite experiment	Four universities	Direct, reflective and hybrid feedback differ in effect	Single study	Design dependence (Section 4.1)
[35]	Survey (n = 2240)	China	Use profiles and institutional support associated with outcomes	Cross-sectional; convenience sample; one university	Social pillar; conditions (Section 4.2)

Table A2. *Cont.*

Ref.	Design	Context	Finding Used in This Review	Main Limitation	Relevance to the SGPF
[38]	Quasi-experimental mixed methods ($n = 176$)	Taiwan	Human–AI contrasts with guided reflection associated with self-directed learning	Non-randomized; one course	Human agency (Section 5)
[42]	Mixed methods ($n = 194$)	Australia	User profiles; AI literacy associated with constructive use	Correlational; self-report	Social pillar (Section 6)
[43]	Survey ($n = 460$)	China	Usefulness, ease of use and norms shape intention to use	Intention rather than behavior; cross-sectional	Social pillar (Section 6)
[45]	Survey, three institutions	Europe	Student and teacher perspectives on ChatGPT	Self-report	European evidence (Section 7)
[46]	Dual-institution study	Sweden	Patterns of student use	Two institutions	European evidence (Section 7)
[47]	Applied governance study	Italy	Applicability of an AI Act governance checklist for assessment	One national context	Institutional pillar (Section 9.1)
[48]	Natural experiment	Norway	Assessment validity under GenAI availability	One setting	Assessment validity (Section 8.2)
[49]	Survey	Slovakia	Digital competences in one department	One department	Policy context (Section 10)
[50]	Tool evaluation (14 detectors)	Europe	Detectors neither accurate nor reliable	Tools change rapidly	Integrity; social pillar (Section 8.1)
[51]	Survey	Cyprus, Greece	Digital technology adoption among marketing communication professionals	Professional, not student, sample	Graduate competence (Sections 8.2 and 16)
[52]	Qualitative study	Europe	Human-centered AI in entrepreneurial ventures	Qualitative; professional context	Graduate competence (Section 16)
[55]	Experimental evaluation	International	Simple adversarial techniques reduce detector accuracy	Specific tools and versions	Integrity (Section 8.1)
[56]	Experimental audit	USA	More than half of non-native writers' essays misclassified as AI-generated	Specific detectors at one point in time	Social pillar (Section 8.1)
[58]	Delphi study	Global	Governance concerns and priorities	Expert opinion	Institutional pillar (Section 9.1)
[60]	Document analysis	South Africa	Progress of documented AI policies can be tracked	Non-European system	Institutional pillar (Section 9.1)
[74]	Qualitative study	China	Personalization valued alongside privacy concerns	One discipline (nursing)	AI literacy; Table 6
[75]	Survey	China	AI literacy linked to sustained digital learning	Cross-sectional; self-report	AI literacy (Section 10)

Table A2. *Cont.*

Ref.	Design	Context	Finding Used in This Review	Main Limitation	Relevance to the SGPF
[76]	Survey (n = 437)	Country not stated	Prompt competence and technology fit linked to educational sustainability	Cross-sectional; self-report	AI literacy (Section 10)
[78]	Technical study	USA	Energy cost of training large language models	Training only; pre-generative models	Environmental pillar (Section 11)

References

1. Yusuf, A.; Pervin, N.; Román-González, M.; Noor, N.M. Generative AI in Education and Research: A Systematic Mapping Review. *Rev. Educ.* **2024**, *12*, e3489. [CrossRef]
2. Chen, S.; Cheung, A.C.K. Effect of Generative Artificial Intelligence on University Students' Learning Outcomes: A Systematic Review and Meta-Analysis. *Educ. Res. Rev.* **2025**, *49*, 100737. [CrossRef]
3. UNESCO. *Education for Sustainable Development: A Roadmap (ESD for 2030)*; UNESCO: Paris, France, 2020.
4. Bianchi, G.; Pisiotis, U.; Cabrera Giraldez, M. *GreenComp: The European Sustainability Competence Framework*; Publications Office of the European Union: Luxembourg, 2022.
5. Boud, D. Sustainable Assessment: Rethinking Assessment for the Learning Society. *Stud. Contin. Educ.* **2000**, *22*, 151–167. [CrossRef]
6. Sejdiu, N.P.; Grassini, S.; Sejdiu, S. Generative AI in Education: Process-Aware Pedagogy, Assessment Integrity, and Institutional Governance. *Open Res. Eur.* **2026**, *6*, 16. [CrossRef] [PubMed]
7. Kuang, W.; Gao, X. Multidimensional Responses to Generative AI in Higher Education: An Integrative Framework Linking Institutional Governance, Teaching Practice, and Student Psychological Adaptation. *Front. Psychol.* **2026**, *17*, 1789506. [CrossRef] [PubMed]
8. Torres-Rivera, A.; Díaz-Torres, S.T.; Díaz-Torres, L.A. Ethical Integration of AI and STEAM Pedagogies in Higher Education: A Sustainable Learning Model for Society 5.0. *Sustainability* **2025**, *17*, 8525. [CrossRef]
9. Sakka, G.; Kavarnou, D.; Anastasou, M.; Karamanidou, M.; Panayiotou, I.; Howieson, W. Institutionalising Sustainability in Higher Education: From Fragmented Initiatives to Embedded Strategic Capability. *Sustainability* **2026**, *18*, 6150. [CrossRef]
10. Miao, F.; Holmes, W. *Guidance for Generative AI in Education and Research*; UNESCO: Paris, France, 2023.
11. European Commission, Directorate-General for Education, Youth, Sport and Culture. *Ethical Guidelines on the Use of Artificial Intelligence (AI) and Data in Teaching and Learning for Educators*; Publications Office of the European Union: Luxembourg, 2022. [CrossRef] [PubMed]
12. Redecker, C. *European Framework for the Digital Competence of Educators: DigCompEdu*; Punie, Y., Ed.; Publications Office of the European Union: Luxembourg, 2017. [CrossRef]
13. Jisc. *AI Maturity Toolkit for Tertiary Education*; Jisc: Bristol, UK, 2024. Available online: https://www.jisc.ac.uk/ai-maturity-toolkit-for-tertiary-education (accessed on 3 September 2026).
14. Whittemore, R.; Knafl, K. The Integrative Review: Updated Methodology. *J. Adv. Nurs.* **2005**, *52*, 546–553. [CrossRef] [PubMed]
15. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
16. Shea, B.J.; Reeves, B.C.; Wells, G.; Thuku, M.; Hamel, C.; Moran, J.; Moher, D.; Tugwell, P.; Welch, V.; Kristjansson, E.; et al. AMSTAR 2: A Critical Appraisal Tool for Systematic Reviews That Include Randomised or Non-Randomised Studies of Healthcare Interventions, or Both. *BMJ* **2017**, *358*, j4008. [CrossRef] [PubMed]
17. Kraft, M.A. Interpreting Effect Sizes of Education Interventions. *Educ. Res.* **2020**, *49*, 241–253. [CrossRef]
18. Cheung, A.C.K.; Slavin, R.E. How Methodological Features Affect Effect Sizes in Education. *Educ. Res.* **2016**, *45*, 283–292. [CrossRef]
19. Egger, M.; Davey Smith, G.; Schneider, M.; Minder, C. Bias in Meta-Analysis Detected by a Simple, Graphical Test. *BMJ* **1997**, *315*, 629–634. [CrossRef] [PubMed]
20. Qian, Y.-F. Pedagogical Applications of Generative AI in Higher Education: A Systematic Review of the Field. *TechTrends* **2025**, *69*, 1105–1120. [CrossRef]
21. Fan, C.-W.; Ke, L.; Chen, Z.-X.; Lv, P. Exploring the Effect of GenAI on Learning Outcomes in Higher Education: A Three-Level Meta-Analysis. *Front. Psychol.* **2026**, *17*, 1758670. [CrossRef] [PubMed]

22. Baig, M.I.; Yadegaridehkordi, E. ChatGPT in the Higher Education: A Systematic Literature Review and Research Challenges. *Int. J. Educ. Res.* **2024**, *127*, 102411. [CrossRef]
23. Liang, Z.; Yang, K.; Sha, L.; Gašević, D.; Yan, L.; Chen, G. A Systematic Review of Generative AI in Education: Empirical Insights from a Human–AI Interaction Perspective. *Br. J. Educ. Technol.* **2026**, *57*, 1221–1271. [CrossRef]
24. Bittle, K.; El-Gayar, O. Generative AI and Academic Integrity in Higher Education: A Systematic Review and Research Agenda. *Information* **2025**, *16*, 296. [CrossRef]
25. Radovan, M.; Košmerl, T.; Radovan, D.M. Environmental Impacts of Generative AI in Education: A Systematic Review of Educational and Technical Evidence. *Sustainability* **2026**, *18*, 7213. [CrossRef]
26. Cong-Lem, N.; Fan, H. Exploring the Effects of Generative AI (GenAI) on Learning Outcomes: A Scoping Review. *Res. Pract. Technol. Enhanc. Learn.* **2026**, *22*, 22. [CrossRef]
27. Wang, G.; Wang, W.; Yang, D.; Ren, J. Generative AI, Cognitive Offloading, and Learner Agency in Higher Education: A Scoping Review. *Behav. Sci.* **2026**, *16*, 1150. [CrossRef] [PubMed]
28. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education: Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
29. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
30. Ateş, H. Human-Centered GenAI Feedback Design in Higher Education: A Multisite Experiment on Direct, Reflective, and Hybrid Approaches to Scientific Argumentation. *Int. J. Educ. Technol. High. Educ.* **2026**, *23*, 38. [CrossRef]
31. Giannakos, M.; Azevedo, R.; Brusilovsky, P.; Cukurova, M.; Dimitriadis, Y.; Hernández-Leo, D.; Järvelä, S.; Mavrikis, M.; Rienties, B. The Promise and Challenges of Generative AI in Education. *Behav. Inf. Technol.* **2025**, *44*, 2518–2544. [CrossRef]
32. Han, X.; Peng, H.; Liu, M. The Impact of GenAI on Learning Outcomes: A Systematic Review and Meta-Analysis of Experimental Studies. *Educ. Res. Rev.* **2025**, *48*, 100714. [CrossRef]
33. Gu, J.; Yan, Z. Effects of GenAI Interventions on Student Academic Performance: A Meta-Analysis. *J. Educ. Comput. Res.* **2025**, *63*, 1460–1492. [CrossRef]
34. Flavell, J.H. Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. *Am. Psychol.* **1979**, *34*, 906–911. [CrossRef]
35. Zhang, J.; Wang, C.; Liu, M. Generative AI Use Profiles and Academic Outcomes in Higher Education: The Role of Institutional Support. *Interact. Learn. Environ.* **2026**. [CrossRef]
36. Felix, S.M.; Lønnum, M.; Lykknes, A.; Staberg, R.L. Teachers' Understanding of and Practices in Critical Thinking in the Context of Education for Sustainable Development: A Systematic Review. *Educ. Sci.* **2025**, *15*, 824. [CrossRef]
37. Risko, E.F.; Gilbert, S.J. Cognitive Offloading. *Trends Cogn. Sci.* **2016**, *20*, 676–688. [CrossRef] [PubMed]
38. Jou, M.; Hao, Y. Fostering Career Awareness and Self-Directed Learning through an AI-Induced Disruptive Learning Environment. *Comput. Educ.* **2026**, *254*, 105688. [CrossRef]
39. Sharma, S. Educational Integrity in GenAI-Augmented Assessment: Making Judgement Visible. *Int. J. Educ. Integr.* **2026**, *22*, 6. [CrossRef]
40. Saito, T. Redefining Agency: A Capability-Driven Research Agenda for Generative AI in Education. *Educ. Sci.* **2026**, *16*, 155. [CrossRef]
41. van Dijk, J.A.G.M. *The Deepening Divide: Inequality in the Information Society*; Sage: Thousand Oaks, CA, USA, 2005.
42. Beckman, K.; Apps, T.; Howard, S.K.; Rogerson, C.; Rogerson, A.; Tondeur, J. The GenAI Divide among University Students: A Call for Action. *Internet High. Educ.* **2025**, *67*, 101036. [CrossRef]
43. Yu, X.; Hu, Y.; Hu, Y.; Li, K. Exploring Factors That Influence Graduate Students' Intention to Use Generative Artificial Intelligence in Academic Research: PLS-SEM and fsQCA Methods. *Educ. Inf. Technol.* **2025**, *30*, 19447–19472. [CrossRef]
44. Gide, E.; Al-Sabbagh, A. Systematised Evidence Mapping of Generative Artificial Intelligence (GenAI) and Digital Divide Phenomena in Higher Education. *Discov. Comput.* **2026**, *29*, 157. [CrossRef]
45. Žáková, K.; Urbano, D.; Cruz-Correia, R.; Guzmán, J.L.; Matišák, J. Exploring Student and Teacher Perspectives on ChatGPT's Impact in Higher Education. *Educ. Inf. Technol.* **2025**, *30*, 649–692. [CrossRef]
46. Nygren, Å.; Eriksson, A.-L.; Sjöberg, J.; Alégroth, E. Exploring Student Use of Generative AI in Higher Education: A Dual-Institutional Study. *Educ. Inf. Technol.* **2026**, *31*, 8703–8730. [CrossRef]
47. Manganello, F.; Nico, A.; Ragusa, M.; Boccuzzi, G. Testing the Applicability of a Governance Checklist for High-Risk AI-Based Learning Outcome Assessment in Italian Universities under the EU AI Act Annex III. *Front. Artif. Intell.* **2025**, *8*, 1718613. [CrossRef] [PubMed]
48. Bratli, H.; Utne, A.; Lynch, M. Assessment Validity in the Age of Generative AI: A Natural Experiment. *Informatics* **2026**, *13*, 56. [CrossRef]

49. Marasová, D.; Kubala, D.; Moravič, M.; Cagaňová, D. A Study on Education in the Department of Transport in Relation to Digital Competences and Skills. *Transp. Logist. Int. J.* **2024**, *24*, 69–78. [CrossRef]
50. Weber-Wulff, D.; Anohina-Naumeca, A.; Bjelobaba, S.; Foltýnek, T.; Guerrero-Dib, J.; Popoola, O.; Šigut, P.; Waddington, L. Testing of Detection Tools for AI-Generated Text. *Int. J. Educ. Integr.* **2023**, *19*, 26. [CrossRef]
51. Komodromos, M.; Harkiolakis, T.; Khoshtaria, T.; Vassiliou, M. Adoption of Digital Technology and Investigation of the Impact in the Marketing Industry for Marketing Communication Professionals: The Case Study of Cyprus and Greece. *Int. J. Technol. Enhanc. Learn.* **2024**, *16*, 173–191. [CrossRef]
52. Komodromos, M. Human-Centred AI: Qualitative Insights, Personalisation, and Customer Experience in European Entrepreneurial Ventures. In *AI-Powered Entrepreneurial Marketing and Communication*; Masouras, A., Maurya, P., Symeou, P., Eds.; IGI Global Scientific Publishing: Hershey, PA, USA, 2026; pp. 145–176. [CrossRef]
53. Scierri, I.D.M.; Picasso, F.; Dicataldo, M.C.; Dipace, A. Institutional Policy Responses to Generative AI in European Higher Education: A Scoping Review of Published Research. *Res. Educ. Media* **2026**, *18*, 10–21. [CrossRef]
54. Peterson, S. Addressing Student Use of Generative AI in Schools and Universities through Academic Integrity Reporting. *Front. Educ.* **2025**, *10*, 1610836. [CrossRef]
55. Perkins, M.; Roe, J.; Vu, B.H.; Postma, D.; Hickerson, D.; McGaughran, J.; Khuat, H.Q. Simple Techniques to Bypass GenAI Text Detectors: Implications for Inclusive Education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 53. [CrossRef]
56. Liang, W.; Yuksekgonul, M.; Mao, Y.; Wu, E.; Zou, J. GPT Detectors Are Biased against Non-Native English Writers. *Patterns* **2023**, *4*, 100779. [CrossRef] [PubMed]
57. European Parliament; Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, L2024/1689. Amended by Regulation (EU) 2026/1744 (Digital Omnibus on AI), Off. J. Eur. Union 2026, L2026/1744. Available online: http://data.europa.eu/eli/reg/2024/1689/oj (accessed on 3 September 2026).
58. Crompton, H.; Burke, D.; Nickel, C.; Bozkurt, A.; Miao, F.; Sharples, M.; Greene, J.A.; Parsons, D.; Gill-Simmen, L.; Edmett, A.; et al. Governing Generative AI in Higher Education: A Global Delphi Study on Policy and Practice. *Int. J. Educ. Technol. High. Educ.* **2026**, *23*, 21. [CrossRef]
59. Kim, B.; Jeong, S.; Cho, B.-K.; Chung, J.-B. AI Governance in the Context of the EU AI Act. *IEEE Access* **2025**, *13*, 144126–144142. [CrossRef]
60. Jacobs-Basadien, M.; Pather, S. Assessing AI Policy and Guideline Progress within the South African Public Higher Education System: Insights from Documented Policies. *Univers. Access Inf. Soc.* **2026**, *25*, 71. [CrossRef]
61. Selwyn, N. *Should Robots Replace Teachers? AI and the Future of Education*; Polity: Cambridge, UK, 2019.
62. Williamson, B.; Eynon, R. Historical Threads, Missing Links, and Future Directions in AI in Education. *Learn. Media Technol.* **2020**, *45*, 223–235. [CrossRef]
63. Williamson, B.; Macgilchrist, F.; Potter, J. Re-Examining AI, Automation and Datafication in Education. *Learn. Media Technol.* **2023**, *48*, 1–5. [CrossRef]
64. Komljenovic, J. The Rise of Education Rentiers: Digital Platforms, Digital Data and Rents. *Learn. Media Technol.* **2021**, *46*, 320–332. [CrossRef]
65. Vakulyk, I.; Voskoboinikova, H.; Haharin, M.; Chabaiovska, M.; Dovzhuk, V.; Tomashevska, M. Globalization of the Education Space in EU Countries and Great Britain through Radical Information Reform. *Amazon. Investig.* **2023**, *12*, 29–38. [CrossRef]
66. European Parliament; Council of the European Union. Regulation (EU) 2016/679 of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *L119*, 1–88.
67. European Commission. *Digital Education Action Plan 2021–2027: Resetting Education and Training for the Digital Age*; COM(2020) 624 Final; European Commission: Brussels, Belgium, 2020.
68. Council of the European Union. Council Recommendation of 23 November 2023 on the Key Enabling Factors for Successful Digital Education and Training. *Off. J. Eur. Union* **2024**, C/2024/1115. Available online: http://data.europa.eu/eli/C/2024/1115/oj (accessed on 3 September 2026).
69. Vuorikari, R.; Kluzer, S.; Punie, Y. *DigComp 2.2: The Digital Competence Framework for Citizens. With New Examples of Knowledge, Skills and Attitudes*; Publications Office of the European Union: Luxembourg, 2022. [CrossRef] [PubMed]
70. European Commission. *Coordinated Plan on Artificial Intelligence 2021 Review*; COM(2021) 205 Final; European Commission: Brussels, Belgium, 2021.
71. EHEA Ministerial Conference. *Rome Ministerial Communiqué*; European Higher Education Area (EHEA): Rome, Italy, 2020.
72. EHEA Ministerial Conference. *Tirana Communiqué*; European Higher Education Area (EHEA): Tirana, Albania, 2024.
73. ENQA; ESU; EUA; EURASHE. *Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG)*; EURASHE: Brussels, Belgium, 2015.

74. Shen, M.; Shen, Y.; Liu, F.-C.; Jin, J.-W. Prompts, Privacy, and Personalized Learning: Integrating AI into Nursing Education. A Qualitative Study. *BMC Nurs*. **2025**, *24*, 470. [CrossRef] [PubMed]
75. Lin, X.; Xu, G.; Xiong, B. Artificial Intelligence Literacy, Sustainability of Digital Learning and Practice Achievement: A Study of Vocational College Students. *PLoS ONE* **2025**, *20*, e0332175. [CrossRef] [PubMed]
76. Gibreel, O.A.; Karataş, K.; Arpaci, I. Prompt Engineering Competence, Knowledge Management, and Technology Fit as Drivers of Educational Sustainability through Generative AI. *Sci. Rep*. **2026**, *16*, 15596. [CrossRef] [PubMed]
77. Miao, F.; Cukurova, M. *AI Competency Framework for Teachers*; UNESCO: Paris, France, 2024.
78. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 3645–3650. [CrossRef]
79. Cañavate, J.; Martínez-Marroquín, E.; Colom, X. Engineering a Sustainable Future through the Integration of Generative AI in Engineering Education. *Sustainability* **2025**, *17*, 3201. [CrossRef]
80. Peng, J.; Li, Y. Frontiers of Artificial Intelligence for Personalized Learning in Higher Education: A Systematic Review of Leading Articles. *Appl. Sci*. **2025**, *15*, 10096. [CrossRef]
81. Esho, E.O.; Akinyelu, A.A.; Dinis, M.A.P. Sustainable Generative AI and Quantum Computing: Review Assessment on the Environmental Impact of Generative AI and Quantum Technologies. *Front. Sustain*. **2026**, *7*, 1726832. [CrossRef]
82. Elkington, J. *Cannibals with Forks: The Triple Bottom Line of 21st Century Business*; Capstone: Oxford, UK, 1997.
83. Sen, A. *Development as Freedom*; Oxford University Press: Oxford, UK, 1999.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.