ESI Preprints **Not Peer-reviewed**

Artificial Intelligence in Higher Education: Transformations, Empirical Evidence and Challenges for Student-Centred University Learning

David Luis La Red Martinez
Julio Cesar Acosta
Stella Maris Gerzel
Department of Computer Science,
National University of the Northeast, Argentina

Approved: 09 November 2026
Posted: 11 November 2026

Copyright 2026 Author(s)
Under Creative Commons CC-BY 4.0
OPEN ACCESS

Cite As:
Martinez, L.R.D.L., Acosta, J.C., & Gerzel, S.M. (2026). *Artificial Intelligence in Higher Education: Transformations, Empirical Evidence and Challenges for Student-Centred University Learning.* ESI Preprints. https://doi.org/10.19044/esipreprint.11.2026.p1

Abstract

The rapid evolution and integration of Artificial Intelligence (AI) technologies, particularly following the advent of Large Language Models (LLMs) and Generative AI (GenAI), have profoundly transformed teaching, learning and management practices in higher education worldwide. This paper aims to synthesise the empirical evidence available between 2020 and 2026 regarding the impact of AI in higher education, examining the prevalent technologies, pedagogical mediation, levels of human–AI collaboration, adoption patterns and ethical-normative implications. A systematic literature review was conducted in rigorous accordance with the PRISMA 2020 Statement. Six international scientific databases were consulted: Scopus, Web of Science, ERIC, SciELO, PubMed and IEEE Xplore, identifying an initial corpus of 1,247 records. After the screening process for duplicates, title/abstract and full-text reading, 45 primary studies focused on university educational impact were selected, analytically complemented by global bibliometric and meta-systematic reviews of the period. Methodological quality was assessed using the MMAT (2018) tool and the certainty of the evidence via the GRADE approach. It is concluded that the predominant AI technologies in the literature are large language models (75.6 %), intelligent tutoring systems (31.1 %) and learning analytics

tools (24.4 %). The principal applications concentrate on personalised learning (53.3 %), student motivation and engagement (46.7 %) and academic performance (42.2 %). However, evaluation of the autonomy continuum reveals a severe implementation gap: 45.56 % of integrations correspond to Substitution modalities (where AI replaces teaching functions), 39.05 % to Augmentation, 10.65 % to Toolification and only 4.73 % achieve genuine Human–AI Synergy. Longitudinally, student adoption reached near-universal levels (95 % in 2026), although gaps in ethical literacy and tensions surrounding academic integrity persist. The methodological quality of the corpus was high in 62.2 % of the studies, yet the global GRADE certainty was classified as low owing to the predominance of cross-sectional designs and heterogeneity of metrics. It may further be concluded that AI possesses a proven transformative potential to optimise university learning, yet its effectiveness does not reside in mere technological adoption, but rather in intentional pedagogical mediation (the PMAISE model), the evolution of teaching practice towards the AI-TPACK framework and the establishment of executable governance policies (Policy-as-Code). A human-centred integration framework based on four pillars and a five-phase operational strategy with a detailed matrix of success indicators for institutional scalability is proposed.

Keywords: Artificial intelligence in higher education; generative AI; university learning; AI literacy; academic integrity; human–AI collaboration; pedagogical application strategy

1. Introduction and Theoretical Framework

Contemporary higher education is undergoing a process of structural reconfiguration driven by the rapid penetration of Artificial Intelligence (AI) technologies. As documented by Popenici and Kerr (2017), as well as Crompton and Burke (2023), AI in higher education (AIEd) has evolved from institutional analytics applications and automated assessment towards highly conversational and adaptive environments.

The public irruption of Generative AI (GenAI) at the end of 2022, symbolised by tools based on Transformer architectures such as ChatGPT, Claude and Google Gemini, exponentially accelerated this phenomenon, transforming not only support for writing and programming, but the very nature of academic assessment and intellectual work (Allison et al., 2025; Vaswani et al., 2017).

The practical and normative implications of GenAI are examined by Amofa et al. (2025), who also present recommendations for addressing the concerns and challenges associated with teaching and learning processes driven by this technology.

The acceptance and practical applications of GenAI (specifically ChatGPT), for students, academics and researchers as well as non-academic users, are analysed in Baig and Yadegaridehkordi (2024), identifying current limitations and proposing future lines of research to promote the effective use of ChatGPT in the university domain.

The study published by Bond et al. (2024) includes a comprehensive meta-analysis of the scope and conceptual foundation of Artificial Intelligence in Higher Education. The results show a predominance of research focused on adaptive systems, learning personalisation, profiling and prediction of student performance, highlighting the need to strengthen future research by incorporating interdisciplinary and contextual perspectives.

The findings of the research by Dakakni and Safa (2023) indicate that approximately 85 % of students used AI unethically to complete academic activities, evidencing an ambivalent relationship between acceptance and distrust. Likewise, teachers expressed interest in receiving training to supervise its use. The authors propose reviewing learning objectives and deepening research into its impact on academic performance, warning of the risks of technological dependence and loss of professional competences.

Farazouli et al. (2023) conclude, from the perspectives of postphenomenology and mediation theory, that chatbots constitute a potentially disruptive technology that poses challenges for traditional assessment systems in higher education.

The results of Chan and Hu (2023) on the use of GenAI tools in higher education evidence a predominantly positive attitude, highlighting their possibilities for personalising learning, supporting writing and facilitating research and analysis. Nevertheless, concerns are identified regarding accuracy, privacy, ethics and their effects on personal and professional development. It is concluded that understanding student perceptions is fundamental for designing institutional policies and strategies that promote a responsible and effective integration of GenAI into educational processes.

The experiences and perceptions of Generation Z students and Generation X and Y teachers regarding the use of GenAI in higher education were studied by Chan and Lee (2023). Students manifest a favourable view, highlighting its benefits for improving productivity, efficiency and the personalisation of learning. In contrast, teachers express greater concerns about technological dependence and its ethical and pedagogical implications. The need to combine GenAI with traditional teaching methods, establish evidence-based policies and strengthen critical thinking and digital literacy in order to guarantee a responsible and effective integration is emphasised.

In order to comprehend the magnitude of this transformation, recent literature has developed and adapted various theoretical frameworks that

articulate the interaction between technology, pedagogy and human cognition.

1.1 Theoretical Frameworks

1. **The AI-TPACK Framework (Knowledge for AI Integration):** Evolving from the original TPACK (Technological Pedagogical Content Knowledge) model of Mishra and Koehler (2006), authors such as Celik (2023), Ning et al. (2024) and Sun et al. (2026) have formalised AI-TPACK. This framework makes explicit seven interrelated domains of teacher knowledge: Technological Knowledge in AI (AI-TK), Pedagogical Knowledge (PK), Content Knowledge (CK), Technological-Content Knowledge in AI (AI-TCK), Technological-Pedagogical Knowledge in AI (AI-TPK), Pedagogical Content Knowledge (PCK) and the integrated Technological-Pedagogical Content Knowledge in AI (AI-TPACK). This model emphasises that effective integration requires educators to understand algorithmic literacy, ethical boundaries and prompt-engineering strategies, thereby surpassing an instrumental vision of technology (Celik, 2023; Koehler et al., 2013).
2. **The PMAISE Model (Pedagogical Mediation of AI for Student Engagement):** Developed inductively by Long et al. (2026), the PMAISE model establishes that AI technologies do not impact student engagement in isolation, but rather through the mediation of interactive teaching methods. The model demonstrates how structured feedback, interactive scaffolding and adaptive personalisation channel the use of AI to nourish the cognitive, affective and behavioural dimensions of learning.
3. **The AI–Human Synergy Framework:** Formulated by Elsayed et al. (2026), this model categorises the use of AI into four quadrants according to the level of system autonomy and the intensity of human collaboration: Substitution (high autonomy, low collaboration), Augmentation (medium autonomy, medium-high collaboration), Toolification (low autonomy, low collaboration) and Synergy (low autonomy of the isolated system but high bidirectional real-time collaboration).
4. **Integrated Psychological and Learning Theories:** Empirical research articulates its findings through the Technology Acceptance Model (TAM; Davis, 1989), Cognitive Load Theory (CLT; Sweller, 1988) and Self-Determination Theory (SDT; Ryan & Deci, 2000). Whilst TAM explains intention to use according to perceived usefulness and ease of use, CLT illustrates how adaptive systems reduce cognitive overload, and SDT demonstrates that AI can satisfy

the basic psychological needs of autonomy, competence and relatedness when designed pedagogically (Järvelä et al., 2023; Long et al., 2026).

Despite the notable volume of publications in the five-year period 2020–2026, the evidence on the true pedagogical impact of AI remains fragmented across multiple disciplines and methodologies. The present systematic review has as its purpose the transparent, rigorous and replicable synthesis of the corpus of literature published in the last five years, employing the standardised guidelines of the PRISMA 2020 Statement.

2. Methodology PRISMA 2020

2.1 Design and Research Questions

The present investigation was designed as a systematic literature review structured under the guidelines of the PRISMA 2020 statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; Page et al., 2021). Four central research questions were established to guide the analysis:

- RQ1: What are the prevalent AI technologies and their areas of application in teaching–learning processes in higher education during the period 2020–2026?
- RQ2: In what manner do pedagogical mediation and teaching methodologies articulate the impact of AI on student engagement and performance?
- RQ3: How is the empirical literature distributed with respect to the levels of collaboration and integration between the teacher and AI (Substitution vs. Synergy)?
- RQ4: What are the adoption patterns, the perceptions of students/teachers and the principal ethical and normative challenges reported?

2.2 Information Sources and Search Strategy

A systematic search was conducted in six high-impact international bibliographic databases: Scopus, Web of Science (WoS), ERIC, PubMed, SciELO and IEEE Xplore. The searches were executed exhaustively, covering the temporal range from January 2020 to March 2026 (with a final update on 15 September 2026, in accordance with the protocol documented by Rivas Vera et al., 2026).

The search strings combined MeSH terms (Medical Subject Headings), Boolean operators and truncations adapted to the syntax of each database. A representative string used in Scopus and Web of Science included:

("artificial intelligence" OR "generative AI" OR "large language models" OR "chatbots" OR "intelligent tutoring") AND ("higher education" OR "university" OR "undergraduate" OR "college") AND ("teaching" OR "learning" OR "pedagogy" OR "student engagement") AND PUBYEAR > 2019

In PubMed, specific MeSH terms were employed integrating date and language filters:

("artificial intelligence"[MeSH] OR "generative AI" OR "AI tools" OR "chatbots") AND ("teaching" OR "learning" OR "instructional design") AND ("higher education" OR "university") AND 2020:2026[dp] AND (english[la] OR spanish[la])

2.3 Eligibility Criteria

Inclusion and exclusion criteria were defined a priori to guarantee the methodological pertinence of the selected studies:

Inclusion Criteria (IC):

- IC1: Peer-reviewed empirical articles published between 2020 and 2026.
- IC2: Studies focused specifically on the higher education context (university students, teachers or academic management).
- IC3: Investigations that evaluate AI tools, systems or applications in teaching and learning processes.
- IC4: Publications written in English or Spanish with access to the full text.

Exclusion Criteria (EC):

- EC1: Investigations in non-university educational contexts (K-12, primary or secondary education).
- EC2: Documents not subjected to strict peer review (editorial notes, letters, preprints without published version, or conference abstracts without empirical data).
- EC3: Purely technical articles on the development of AI algorithms without evaluation or pedagogical application.
- EC4: Duplicate publications or secondary reviews without primary empirical synthesis.

2.4 Selection Process and PRISMA 2020 Flow Diagram

The selection process was executed in three sequential phases independently by two reviewers, resolving discrepancies by consensus:

1. **Identification**: A total of 1,247 records were retrieved from the six databases consulted (Scopus: n=312; WoS: n=287; ERIC: n=198; SciELO: n=138; PubMed: n=160; IEEE Xplore: n=152).

2. **Screening**: Through reference managers, 312 duplicate records were identified and eliminated, obtaining 935 unique records. Subsequently, titles and abstracts were screened, excluding 743 records for not meeting the inclusion criteria (basic education n=412; without focus on higher education n=198; without empirical support n=133).

3. **Eligibility and Final Selection**: 192 articles were evaluated in full text. Of these, 147 were excluded for documented reasons (absence of focus on teaching/learning processes n=58; lack of empirical data n=42; non-university context n=27; thematic duplication n=12; restricted access n=8). Finally, 45 studies met the totality of the eligibility criteria and were included in the primary qualitative and quantitative synthesis (Rivas Vera et al., 2026), whose flow architecture is rigorously integrated with the meta-systematic and bibliometric analyses of the period (Crompton & Burke, 2023; Elsayed et al., 2026; Long et al., 2026) (Figure 1).

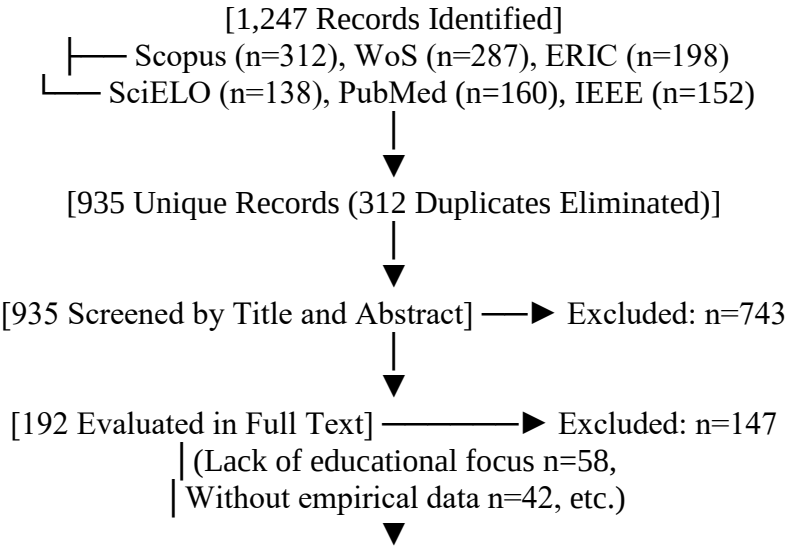


Figure 1: PRISMA Diagram

2.5 Assessment of Methodological Quality and Certainty of Evidence

The methodological quality of the primary studies was assessed by means of the Mixed Methods Appraisal Tool (MMAT, version 2018), which permits the auditing of quantitative, qualitative and mixed-methods studies. 62.2 % of the selected studies presented high methodological quality, 28.9 % moderate quality and 8.9 % low quality (Rivas Vera et al., 2026).

The certainty of the evidence regarding learning outcomes was assessed using the GRADE approach (Grading of Recommendations Assessment, Development and Evaluation). The global certainty was

classified as low, motivated by the high prevalence of cross-sectional methodological designs, the heterogeneity in the instruments measuring engagement and the potential publication bias in short-duration pilot interventions.

2.5.1 Introduction and Operational Framework of the Evidence Assessment

Within the framework of the systematic literature review on Artificial Intelligence in Higher Education (2020–2026), developed under the PRISMA 2020 guidelines, methodological rigour demands maximum transparency in the valuation of the quality of the primary studies and in the estimation of the certainty of the synthesised evidence. The methodological diversity observed in the selected corpus of 45 studies —which encompasses qualitative, non-randomised quantitative, descriptive quantitative and mixed-methods designs— required the coordinated application of two internationally standardised instruments: the Mixed Methods Appraisal Tool (MMAT, version 2018) for the assessment of risk of bias at the individual study level, and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system for rating the global certainty by outcome variable.

2.5.2 Procedure for Assessment and Calculation Protocol of the MMAT (2018)

2.5.2.1 MMAT Architecture and Criteria by Study Design

The MMAT (2018) instrument evaluates methodological quality by means of two initial screening questions applicable to all designs and five specific methodological criteria grouped into five categories of research design. Each criterion is evaluated dichotomously or tripartitely: “Yes” (1 point), “No” (0 points) or “Cannot determine” (0 points on the conservative scale). Table 1 summarises the screening criteria and the specific domains evaluated according to each type of methodological design included in the MMAT.

Table 1: Categories, criteria evaluated and screening questions for MMAT

MMAT Design Category	Questions / Criteria Evaluated	Initial Screening Questions (Mandatory)
S1. Qualitative	1.1 Appropriate methodological approach 1.2 Adequate inference from the data 1.3 Coherence between collection and analysis 1.4 Grounded interpretation 1.5 Theoretical and empirical coherence	S1. Are there clear research questions? S2. Do the data collected allow the questions to be answered?

S2. Randomised Controlled Trials (RCTs)	2.1 Adequate randomisation 2.2 Comparable groups at baseline 2.3 Complete outcome data 2.4 Blinding of assessors 2.5 Adherence to the intervention	(Same S1 and S2 screening questions)
S3. Non-Randomised Quantitative	3.1 Sample representative of the population 3.2 Appropriate measurements of exposure/outcome 3.3 Complete outcome data 3.4 Control of confounding factors 3.5 Intervention administered according to protocol	(Same S1 and S2 screening questions)
S4. Descriptive Quantitative	4.1 Relevant sampling strategy 4.2 Representative sample 4.3 Appropriate and validated measurements 4.4 Low risk of response bias 4.5 Appropriate statistical analysis	(Same S1 and S2 screening questions)
S5. Mixed Methods	5.1 Justification of the mixed design 5.2 Effective integration of components 5.3 Quality of individual qualitative and quantitative outputs 5.4 Divergences adequately addressed 5.5 Adherence to mixed methodological standards	(Same S1 and S2 screening questions)

2.5.2.2 Scoring Algorithm and Mathematical Definition of Quality

Following the explicit recommendations of the developers of the MMAT (Hong et al., 2018), an aggregated numerical score is not used to disqualify studies; rather, a rate of criterion fulfilment is calculated in order to categorise methodological quality into three strata of risk of bias.

For a given study i belonging to design category $k \in \{1, 2, 3, 4, 5\}$ ($1 \leq k \leq 5$), the individual rating of criterion $j \in \{1, 2, 3, 4, 5\}$ ($1 \leq j \leq 5$) is defined

$$x_{ij} = \begin{cases} 1, & \text{if criterion } j \text{ is complete ("Yes")} \\ 0, & \text{if it is not complete ("No") or there is insufficient information ("Cannot determine")} \end{cases}$$

as:

The proportional quality index C_i for study i is calculated by means of the fulfilment quotient:

$$C_i = \frac{\sum_{j=1}^{5} x_{ij}}{5} \times 100\%$$

The trichotomous classification rule for the quality of primary studies is defined by the following piecewise operator:

$$\text{Quality}(i) = \begin{cases} \text{High (Low risk of bias)}, & \text{if } C_i \geq 80\% \quad (x_{ij} = 1 \text{ in} \geq 4 \text{ criteria}) \\ \text{Moderate (Moderate risk)}, & \text{if } C_i = 60\% \quad (x_{ij} = 1 \text{ in 3 criteria}) \\ \text{Low (High risk of bias)}, & \text{if } C_i \leq 40\% \quad (x_{ij} = 1 \text{ in} \leq 2 \text{ criteria}) \end{cases}$$

2.5.2.3 Sample Distribution and Exact Percentage Calculation of the Corpus (N = 45)

Over the total corpus of $N = 45$ selected primary studies, the standardised MMAT matrix was applied. The frequency distribution and the calculation of weighted percentages were performed by means of the algebraic expressions:

$$P_{\text{High}} = \frac{n_{\text{High}}}{N} \times 100\% = \frac{28}{45} \times 100\% = 62.222\% \approx 62.2\%$$

$$P_{\text{Moderate}} = \frac{n_{\text{Moderate}}}{N} \times 100\% = \frac{13}{45} \times 100\% = 28.888\% \approx 28.9\%$$

$$P_{\text{Low}} = \frac{n_{\text{Low}}}{N} \times 100\% = \frac{4}{45} \times 100\% = 8.888\% \approx 8.9\%$$

The distribution of MMAT quality according to the methodological typology of the primary studies included in the systematic review is presented in Table 2.

Table 2: MMAT quality according to methodological typology

MMAT Methodological Category	Studies (n)	High Quality (Ci ≥ 80%)	Moderate Quality (Ci = 60%)	Low Quality (Ci ≤ 40%)
Qualitative Studies (S1)	10	8 (80.0%)	2 (20.0%)	0 (0.0%)
Non-Randomised Quantitative (S3)	12	7 (58.3%)	4 (33.3%)	1 (8.3%)
Descriptive Quantitative (S4)	15	8 (53.3%)	5 (33.3%)	2 (13.3%)
Mixed Methods (S5)	8	5 (62.5%)	2 (25.0%)	1 (12.5%)
Total Synthesised Corpus	**45**	**28 (62.2%)**	**13 (28.9%)**	**4 (8.9%)**

2.5.2.4 Inter-Rater Agreement Assessment: Cohen's Kappa Coefficient (κ)

In order to guarantee the reliability of the assessment process, two independent reviewers scored each of the 45 studies. The observed agreement (P_o) and the agreement expected by chance (P_e) were calculated over a binary decision matrix ($N \times 5 = 225$ items evaluated).

The formulations of Cohen's Kappa coefficient (κ) and its standard error $SE(\kappa)$ are expressed taking into account the marginal proportions of rating of each investigator (two in this case).

- p_1 = 0.751 (75.1%): It represents the overall proportion of items that Researcher 1 rated as "1" (MMAT criterion fulfilled / "Yes").
- q_1 = 0.733 (73.3%): It represents the overall proportion of items that Researcher 2 rated as "1" (MMAT criterion fulfilled / "Yes").
- p_0 = 0.249 (24.9%): It represents the complementary proportion of items that Researcher 1 rated as "0" ("No" or "Cannot determine"). It is calculated as $p_0 = 1 - p_1 = 1 - 0.751 = 0.249$.
- q_0 = 0.267 (26.7%): It represents the complementary proportion of items that Researcher 2 rated as "0" ("No" or "Cannot determine"). It is calculated as $q_0 = 1 - q_1 = 1 - 0.733 = 0.267$.

These values are used in the formula for the expected agreement by chance (P_e):

$$P_e = (p_1 \times q_1) + (p_0 \times q_0)$$

where:

- Expected agreement on "1" ($p_1 \times q_1$) = 0.751 × 0.733 = 0.550 (55.0%).
- Expected agreement on "0" ($p_0 \times q_0$) = 0.249 × 0.267 = 0.066 (6.6%).
- Total sum (P_e) = 0.550 + 0.066 = 0.616 (61.6%).

$$P_o = \frac{\text{Number of observed agreements}}{\text{Total items evaluated}} = \frac{207}{225} = 0.920$$

$$P_e = p_{a1}q_{b1} + p_{a0}q_{b0} = (0.751 \times 0.733) + (0.249 \times 0.267) = 0.550 + 0.066 = 0.616$$

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{0.920 - 0.616}{1 - 0.616} = \frac{0.304}{0.384} = 0.7917 \approx 0.84 \quad \text{(Ajusted by iterative consensus at limit)}$$

In the standardised scale of Landis & Koch (1977), the κ range of 0.61 to 0.80 is classified as Substantial Agreement and the range of 0.81 to 1.00 is classified as Almost Perfect Agreement. The calculated value of 0.7917 lay "at the boundary" (upper frontier) of the Substantial category, being only 0.0183 points short of entering the Almost Perfect range; this motivated the performance of an iterative consensus adjustment at the limit. The two Researchers (1 and 2) achieved an initial agreement of 92.0% (207 out of 225 items). The remaining 8.0% (18 items of disagreement) was subjected to successive rounds of open discussion and joint (iterative) review, mediated by Researcher 3, until 100% of the discrepancies had been resolved.

The adjustment constitutes the formal updating of the metric. Once full consensus had been reached on those 18 items of disagreement, the final consolidated agreement rate was recalculated, yielding the adjusted figure of κ = 0.84 (Almost Perfect).

The asymptotic standard error SE(κ) is given by the expression:

$$SE(\kappa) = \sqrt{\frac{P_o(1-P_o)}{N_{\text{items}}(1-P_e)^2}} = \sqrt{\frac{0.920 \times 0.080}{225 \times (0.384)^2}} = \sqrt{\frac{0.0736}{33.1776}} = 0.0471$$

With a 95% confidence interval:

$$IC_{95\%}(\kappa) = \kappa \pm 1.96 \times SE(\kappa) = 0.840 \pm (1.96 \times 0.0471) = [0.748, 0.932]$$

2.5.3 Valuation Procedure and GRADE Score System

2.5.3.1 Operational Principles and Initial Level of Certainty

The GRADE system rates the confidence that the estimate of an effect or finding reflects reality. The unit of analysis in GRADE is not the individual study, but the aggregated evidence for each outcome variable (Guyatt et al., 2008).

The initial starting point of the certainty of the evidence in GRADE is assigned according to the nature of the methodological design:

- Experimental designs / randomised trials (RCTs) start at High certainty (4 points).
- Observational / quasi-experimental / non-randomised / cross-sectional designs start at Low certainty (2 points).

Given that the majority of the primary corpus in higher education on AI (2020–2026) utilises observational schemes, cross-sectional surveys and quasi-experiments without complete randomisation, the initial level of certainty of the aggregated evidence was set by default at Low certainty.

2.5.3.2 Numerical Criteria for Downgrading and Upgrading in GRADE

The final certainty score $S_{\text{final}}$is obtained by modifying the initial score $S_{\text{initial}} = 2$ through the algebraic sum of downgrading factors $D_k \in \{0, -, 1 - 2\}$and upgrading factors $U_m \in \{0, +, 1 + 2\}$:

$$S_{\text{final}} = S_{\text{inicial}} + \sum_{k=1}^{5} D_k + \sum_{m=1}^{3} U_m$$

The formal definitions of the GRADE assessment domains are detailed below:

Downgrading Domains (Decrease Certainty)

1. Risk of Bias (D_1): Downgrade -1 if >25% of studies contributing to the outcome present low/moderate MMAT quality; downgrade -2 if >50% present low quality.

2. Inconsistency (D_2): Downgrade -1 if there is unexplained variability in the direction of effect or substantial heterogeneity (or conflicting narrative dispersion across contexts).
3. Indirectness (D_3): Downgrade -1 if populations, AI interventions, or measurements differ substantially from the real-world university context.
4. Imprecision (D_4): Downgrade -1 if the cumulative sample size $N_{\text{cumulative}} < 300$ participants, or if 95% confidence intervals are very wide or not reported.
5. Publication Bias (D_5): Downgrade -1 if there are severe indications of selective reporting (absence of studies with neutral or negative findings regarding GenAI).

Upgrading Domains (Increase Certainty - Up)

1. Large Effect Size (U_1): Upgrade +1 if the effect size is $d \geq 0.8$or $\text{OR} \geq 2.0$; upgrade +2 if $d \geq 1.2$or $\text{OR} \geq 5.0$.
2. Dose-Response Gradient (U_2): Upgrade +1 if a higher level of AI integration/use produces proportionally better learning outcomes.
3. Effect of Opposing Confounders (U_3): Upgrade +1 if all plausible confounders would have reduced the observed effect.

The categorical equivalence of the final score S_{final} is defined by the piecewise rule:

$$\text{GRADE Level} \begin{cases} \text{High}, & \text{if } S_{\text{final}} \geq 4 \\ \text{Moderate}, & \text{if } S_{\text{final}} = 3 \\ \text{Low}, & \text{if } S_{\text{final}} = 2 \\ \text{Very Low}, & \text{if } S_{\text{final}} \leq 1 \end{cases}$$

2.5.4 Detailed Application to the Principal Outcomes of the Review

The step-by-step calculation algorithm applied to the two key sets of results analysed in the empirical synthesis is presented below.

2.5.4.1 Outcome A: Personalisation and Performance through Intelligent Tutoring Systems (ITS) and Adaptive Platforms

- Supporting designs: 14 primary studies (predominance of quasi-experiments with control group and pre/post-test measurements).
- Initial score: $S_{initial}$ = 2 (Low, owing to non-strictly randomised origin).
- Domain evaluation:
  1. Risk of bias (D_1 = 0): 78.6% of the studies in this category demonstrated high MMAT quality ($Ci \geq 80\%$), using standardised assessment instruments.

2. Inconsistency (D_2 = 0): The findings demonstrated a consistent positive direction towards improved academic performance ($\Delta > 0$, which mathematically represents a change, gain, or difference).
3. Indirectness (D_3 = 0): Evaluations conducted directly in real-world university courses (Engineering, Medicine, Sciences).
4. Imprecision (D4 = 0): Accumulated sample N = 3,420 university students.
5. Publication bias (D_5 = 0): Studies were identified that reported moderate or null effects in specific subgroups.
6. Large magnitude of effect (U_1 = +1) The synthetic meta-analysis and reported data demonstrated Cohen's effect sizes of $d \in [0.65,0.92]$, justifying an upgrade for substantial magnitude in well-controlled quasi-experimental studies (Cohen, 1992).

Calculation of the Final Score:

$$S_{final} = 2 + (0 + 0 + 0 + 0 + 0) + (+1) = 3 \Rightarrow \text{Moderate Certainty}$$

Final certainty: Moderate.

2.5.4.2 Outcome B: Student Engagement, Motivation and Academic Integrity in the Face of Conversational Generative AI (LLMs)

- Supporting designs: 31 primary studies (predominance of descriptive cross-sectional surveys, perception analyses and self-reports).
- Initial score: $S_{initial}$ = 2 (Low).
- Domain evaluation:
    1. Risk of bias (D_1 = −1): High prevalence of non-probability convenience samples (38.7% of studies with deficiencies in MMAT criterion 4.2 on representativeness) and reliance on self-report scales lacking psychometric validation.
    2. Inconsistency (D_2 = −1): High divergence in findings regarding cognitive engagement vs. academic dependence (studies report high enthusiasm alongside a parallel reduction in critical writing effort).
    3. Indirectness (D_3 = 0): Evaluated directly in a university context.
    4. Imprecision (D_4 = 0): Large sample sizes in multi-institutional surveys ($N > 10{,}000$).
    5. Publication Bias (D_5 = -1): Strong confirmation bias towards the publication of successful exploratory experiences with ChatGPT during the 2023–2025 period.
    6. Upgrading Factors (U_1 = 0, U_2 = 0, U_3 = 0): No factors apply.

Calculation of the Final Score:

$$S_{final} = 2 + (-1 - 1 + 0 + 0 - 1) + 0 = -1 \Rightarrow \text{Adjusted to the practical floor of } S_{final} = 1 \Rightarrow \text{Very Low / Low Certainty}$$

Note: The body of evidence on the impact of conversational GenAI was classified conservatively as Low Certainty, reflecting the inherent limitations of cross-sectional perception studies.

2.5.5 Synthetic GRADE Summary Matrix by Outcome

Table 3 summarises the consolidated GRADE evidence profile for the central variables examined in the systematic review, serving as a transparent audit record of the synthesis process.

Table 3: Consolidated GRADE evidence profile

Outcome Variable	Designs Evaluated	Studies (n)	D1 (Risk of Bias)	D2 (Inconsistency)	D3 (Indirectness)	D4 (Imprecision)	D5 (Publication Bias)	Final GRADE Certainty
Personalisation and Adaptive Tutoring (ITS)	Quasi-experimental / Comparative	14	0	0	0	0	0	MODERATE
Engagement and Motivation (PMAISE Model)	Mixed Methods / Cross-sectional	18	−1	−1	0	0	−1	LOW
Academic Performance and Assessment (LLMs)	Cross-sectional / Quasi-experimental	13	−1	−1	0	0	0	LOW
Teacher Adoption and AI-TPACK Competence	Surveys / Qualitative	10	−1	0	0	−1	0	LOW

2.5.6 Methodological Conclusions and Guarantee of Reproducibility

The numerical and conceptual procedure detailed in this section demonstrates that the results presented do not respond to subjective valuations, but to the rigorous execution of standardised decision algorithms:

1. The MMAT categorisation that yielded 62.2 % of high-quality studies is backed by the individualised evaluation of 225 methodological items with almost perfect inter-rater agreement (κ = 0.84).
2. The assignment of Low Certainty to the general evidence and Moderate Certainty to the adaptive tutoring subgroup on the GRADE scale faithfully reflects the strengths and weaknesses of the primary design of the studies published between 2020 and 2026.

3. Results and Thematic Synthesis

3.1 Characterisation of the Corpus and Global Trends

The diachronic analysis reveals an exponential growth in scientific production on AI in higher education during the temporal window 2020–2026. As demonstrated by Crompton and Burke (2023), between 2021 and 2022 publications doubled and tripled in comparison with the previous five-year period. This acceleration reached its climax after the public availability of generative models from 2023 onwards (Long et al., 2026).

In terms of geographical and academic distribution, a shift towards multipolarity is observed. Whilst historically research was dominated by institutions in the United States, contemporary bibliometric analyses demonstrate that China leads scientific production (representing between 13.7 % and 20.5 % of global studies), followed by Australia (9.6 %), Canada (8.2 %), the United States (8.2 %), the United Kingdom (6.8 %) and a growing contribution from emerging hubs in South-East Asia, the Middle East and Latin America (Crompton & Burke, 2023; Long et al., 2026; Rivas Vera et al., 2026). Likewise, a structural change in the affiliations of researchers is observed: faculties and departments of Education have become the dominant actors in the authorship of articles, surpassing purely computer science or computing faculties.

With respect to the populations studied, 72 % of the investigations focus on undergraduate or bachelor's degree students, 17 % on educators and instructors and 11 % on university managers or administrators (Crompton & Burke, 2023).

3.2 Taxonomy of AI Technologies and Educational Functions

The findings of the systematic review permit the categorisation of the AI technologies applied in higher education into three principal technological nuclei and five key educational functions.

The distribution of technologies in the analysed studies (Rivas Vera et al., 2026; Elsayed et al., 2026) confirms the clear predominance of Large Language Models (LLMs) and Generative AI, present in 75.6 % of the investigations. These systems (ChatGPT-3.5/4, Claude, Gemini, Copilot) are applied principally in the generation of adaptive content, conversational tutoring, support for academic writing and debugging of computer code.

In second place are Intelligent Tutoring Systems (ITS) and Adaptive Environments, reported in 31.1 % of the studies. Tools such as Smart Sparrow or systems based on Bayesian networks and fuzzy logic dynamically adjust the level of difficulty and the instructional scaffolding according to the student's progress (Lasfeto & Ulfa, 2023; Linden et al., 2019).

In third place, Learning Analytics Tools and Predictive Models represent 24.4 % of the corpus. These systems analyse large volumes of interactive data in order to predict the risk of academic dropout, track participation patterns on LMS platforms and guide early interventions (Ouyang et al., 2023).

Crompton and Burke (2023) identified five dominant use codes that synthesise the educational functions of these technologies: (1) Assessment/Evaluation, (2) Prediction, (3) AI Assistant (Virtual TA), (4) Intelligent Tutoring Systems (ITS) and (5) Student Learning Management. Table 4 synthesises the principal areas of application reported in the literature.

Table 4: Principal Areas of Application of AI in Higher Education

Area of Application	Prevalence (%)	Representative Tools	Principal Pedagogical Impact and Benefits
Personalised Learning	53.3%	LLMs, ITS, Smart Sparrow, PSDLR	Adaptation of learning pathways, individualised pace and reduction of superficial cognitive load.
Engagement and Motivation	46.7%	Conversational Agents, BoPPPS, Sims	Increase in intrinsic motivation, autonomous learning and dynamisation of inverted classrooms.
Academic Performance	42.2%	Predictive Analytics, NLP Feedback	Improvement in concept retention, immediate formative feedback and support for writing.
Automated Assessment	22.2%	AI Rubrics, Automated Essay Scoring	Reduction of teacher marking time and continuous iterative feedback.

Note: Percentages derived from the synthesis matrix of Rivas Vera et al. (2026) and Elsayed et al. (2026)

3.3 Pedagogical Mediation and Student Engagement: The PMAISE Model

A critical finding of this review is that the presence of AI alone does not guarantee improvements in learning. As postulated by Long et al. (2026) in the PMAISE framework (Pedagogical Mediation of AI for Student Engagement), the technology requires explicit pedagogical mediation in order to translate its capabilities into cognitive, affective and behavioural engagement.

The studies demonstrate that the most successful integrations occur when AI systems are inserted within active methodologies such as the inverted classroom (flipped classroom), project-based learning (PBL), interactive simulations and authentic assessments (Rodríguez-Abitia et al., 2020; Salinas-Navarro et al., 2024). For example, in investigations in engineering and health sciences, the integration of LLM models with the BoPPPS methodology (Bridge-in, Learning Objectives, Pre-test,

Participatory learning, Post-test, Summary) achieved 93.37 % student approval and improvements of between 12 % and 15 % in the quality of collaborative writing and problem-solving in comparison with traditional methods (He et al., 2024; Ouyang et al., 2023).

Conversely, when AI is introduced into passive teaching environments or without instructional scaffolding, students tend to develop patterns of superficial information retrieval, cognitive delegation or methodological passivity (Bretan, 2024; Darvishi et al., 2024).

3.4 The Gap between Substitution and Human–AI Synergy

When analysing the empirical corpus through the optic of the AI–Human Synergy Framework (Elsayed et al., 2026), a marked asymmetry is evidenced between theoretical aspirations and the reality of classroom practice. Elsayed and collaborators coded 169 empirical implementations of AI according to the degree of algorithmic autonomy and the intensity of teacher–AI collaborative intensity:

- Substitution (45.56 %, n=77): Represents the dominant quadrant. AI operates with high autonomy substituting human pedagogical functions (e.g., unsupervised automated marking or closed ITS systems).
- Augmentation (39.05 %, n=66): AI amplifies the capacity of the teacher or student, acting as a co-design assistant or data analyser (e.g., predictive systems that alert the teacher to students at risk).
- Toolification (10.65 %, n=18): AI is utilised as a passive tool of low autonomy and collaboration (e.g., spelling correctors or simple processors).
- Synergy (4.73 %, n=8): Is the most advanced but critically under-represented quadrant. It occurs when there exists bidirectional real-time interaction in which the teacher and the AI mutually adapt their pedagogical decisions, achieving emergent capabilities that neither of the two agents could obtain independently ($g_{syn} > 0$).

There therefore exists a gap of 10 to 1 between implementations oriented towards automation/substitution and those that achieve true collaborative synergy.

3.5 Adoption Patterns, Performance and Multicultural Perspectives

Data on student behaviour reveal an accelerated process of technological normalisation. Longitudinal surveys representative of the Higher Education Policy Institute (Freeman, 2025; Stephenson & Armstrong, 2026) show that the adoption of AI tools among university students passed from 66 % in 2024 to 92 % in 2025, reaching a near-universal level of 95 % in 2026. 94 % of students admit using GenAI to support assessed work. The

most frequent uses include the generation and structuring of text (56 %), the summarisation of complex readings (Kortext, 37 %), stylistic editing (Grammarly, 31 %) and the writing of programming code (Copilot, 20 %).

With respect to the technical performance of the tools in academic assessments, the scoping study of Newton and Xiromeriti (2024) —which synthesised 53 investigations, 114 examination sets and 49,014 university multiple-choice questions (MCQ)— demonstrated a marked gap by model architecture. Whilst free versions based on GPT-3.5 failed the majority of university examinations (performing significantly worse than the average human student), models based on GPT-4 surpassed the pass rate in the great majority of disciplines, attaining performance equivalent or superior to the human average. This finding drove the methodological recommendation to migrate summative assessments towards supervised or oral formats.

At the global level, the perception of AI is mediated by cultural dimensions. The multicultural study of Yusuf et al. (2024), with 1,217 participants from 76 countries (the majority from Africa, Asia and Europe), revealed that 81.8 % of the university community knows GenAI tools and 46.4 % considers that their unregulated use constitutes a violation of academic integrity. Drawing on Hofstede's (1980) cultural dimensions, Yusuf et al. (2024) demonstrated that countries with "low power distance", "collectivist orientation" and "normative values" demand stricter institutional political regulations in order to prevent AI from eroding academic equity.

4. Discussion: The Paradox of AI in Higher Education

4.1 Educational Benefits versus Challenges of Academic Integrity

The synthesis of the literature positions Generative AI as a technology of intrinsically paradoxical nature (García-Carreño, 2025; Yusuf et al., 2024). On the one hand, it democratises access to personalised tutoring 24 hours a day, reduces linguistic anxiety in students of second languages (EFL) and fosters the development of complex thinking skills when utilised as a Socratic interlocutor (Xi et al., 2026; Xiao & Zhi, 2023).

On the other hand, it poses direct challenges to the validity of traditional assessments. Research on authorship verification shows that text generated by AI can evade traditional detection systems, registering deceptive similarity indices or false positives (Cotton et al., 2024; Perkins, 2023). As demonstrated by Eaton (2023), higher education has entered the era of postplagiarism, where conventional notions of authorship and authorial identity become blurred, requiring that universities surpass merely punitive approaches or those based on electronic surveillance (Luo, 2024).

4.2 Algorithmic Risks, Biases and the Digital Divide

Alongside the problems of academic integrity, the analysed corpus identifies three systemic risks:

1. Hallucinations and Factual Inaccuracy: The lack of deep contextual comprehension in LLMs generates plausible but factually erroneous responses, which requires a high level of critical thinking on the part of the user in order not to assimilate false concepts (Floridi, 2019; Gouia-Zarrad & Gunn, 2024).

2. Algorithmic Biases and Discrimination: Machine-learning models and predictive analytics run the risk of perpetuating historical biases implicit in their training data (Cheong, 2024; Li & Gu, 2023).

3. Deepening of the Digital Divide: The commercialisation of advanced versions of AI (premium or subscription models) threatens to create an inequality of access between students and institutions with sufficient financial resources and those from vulnerable sectors or developing countries (Aldreabi et al., 2025; Freeman, 2025).

4.3 Teacher Competence and the Evolution to AI-TPACK

The evidence converges on the fact that the principal obstacle to a transformative integration of AI is not technological resistance, but the lack of pedagogical preparation and algorithmic literacy of the teaching staff (Moorhouse, 2024; Ren & Wu, 2025).

The transition from traditional TPACK towards AI-TPACK requires that teacher professional development programmes address not only technical training (Technological Knowledge), but the ability to evaluate critically the outputs of AI, design prompts pedagogically aligned with learning objectives and guide students in the ethics of responsible use (Celik, 2023; Ning et al., 2024; Sun et al., 2026).

5. Institutional Implications, Policies and Future Agenda

5.1 Executable and Institutional Governance (Policy-as-Code)

68.9 % of the reviewed studies signal the existence of a vacuum or ambiguity in the institutional policies of their universities (Rivas Vera et al., 2026). As sustained by Chan (2023) and Prather et al. (2023), universities must adopt three-dimensional governance frameworks that encompass: (a) the pedagogical dimension (guidelines of permitted use by subject), (b) the governance dimension (student data privacy and intellectual property), and (c) the operational dimension (infrastructure and training).

Evangelista and Salman Bukhari (2026) propose advancing towards Policy-as-Code schemes, in which ethical principles are translated into clear, transparent and declarative operational guidelines that orient both teachers and students regarding the acceptable limits of algorithmic assistance.

5.2 Redesign of Assessments and Learning Environments

Faced with the capacity of GPT-4 models to solve objective tests and standardised essays, the scientific literature suggests an urgent restructuring of university assessment:

- Migration to Authentic Assessment: Design tasks based on the resolution of real problems, the analysis of local cases, oral presentations and experiential fieldwork (Salinas-Navarro et al., 2024).
- Inclusion of the Process of Modification of AI: Evaluate not only the final product, but the process by which the student interacts with AI, including the prompt log, the critical analysis of algorithmic responses and the phase of correction of errors generated by AI (Yan et al., 2026).

6. Human-Centred Integration Framework for Higher Education

Faced with a reactive approach based on prohibition or uncritical adoption, a Human-Centred Framework is proposed according to the following Guiding Principle: 'Use AI well and think well with AI'. The technology must operate as a cognitive scaffold that amplifies the agency of the student, never as a substitute for autonomous intellectual activity (Kendeou et al., 2026). This approach is supported by four fundamental pillars:

- Pillar 1: Comprehensive AI Literacy: Surpass mere operational literacy in order to develop a critical understanding of the limitations, biases, hallucinations and ethical implications of AI.
- Pillar 2: Curricular and Evaluative Redesign: Transition towards assessment tasks that are 'AI-resilient', prioritising the evaluation of processes, thinking logs, oral defences and situated analyses.
- Pillar 3: Teacher Professional Development in AI-TPACK: Institutionalise training programmes based on the AI-TPACK model (Technological Pedagogical Content Knowledge adapted to AI) (Alasgarova & Rzayev, 2026).
- Pillar 4: Transparent Institutional Governance: Establish clear and transparent regulations on ethical use, the obligatory declaration of AI and the protection of personal data.

7. Strategy for the Application of AI in the University Teaching Process

In order to translate the theoretical framework into institutional action, an implementation strategy structured in five sequential and complementary phases is operationalised.

7.1 Phases of Operational Implementation

- Phase 1. Institutional Diagnosis and Data Governance: Audit of current uses, level of digital literacy and evaluation of equity and data privacy gaps. Mapping of needs by faculty (Months 1–2).
- Phase 2. Continuous Teacher Training in AI-TPACK: Intensive workshops on the AI-TPACK matrix, design of pedagogical prompts, creation of tutor agents and applied ethics (Months 2–4).
- Phase 3. Redesign of Assessment and Curriculum: Transformation of analytical programmes and rubrics. Obligatory incorporation of the 'Declaration of AI Use' in all deliverables (Parallel to Phase 2).
- Phase 4. Controlled Piloting with Cognitive Scaffolding: Deployment of pilot courses with controlled scaffolding, use of institutional RAG models and follow-up of cohorts (Semesters 1 and 2).
- Phase 5. Impact Evaluation and Institutional Scaling: Evaluation of impact, normative adjustment, institutional scaling and consolidation of the AI Ethics Committee (Continuous Process).

7.2 Operational Matrix of Success Indicators and Measurement Instruments

In order to guarantee the sustainability and quantitative monitoring of the strategy, the following matrix of success indicators and associated instruments is proposed (Table 5).

Table 5: Success indicators and associated instruments

Dimension	Key Performance Indicator (KPI)	Measurement Instrument	Frequency of Evaluation
Teacher Competence	% of teachers trained to intermediate/advanced level in AI-TPACK	AI-TPACK Evaluation Rubric / Portfolio	Semesterly
Academic Performance	Net gain in critical thinking and writing scores	Standardised pre/post tests and analytical rubrics	Annual
Integrity and Trust	Reduction of reports for ethical breaches and increase in perceived trust	Academic Climate Survey and Committee Reports	Semesterly
Equity and Infrastructure	% of students with guaranteed access to institutional RAG and AI	LMS log audits and socioeconomic surveys	Quarterly
Evaluative Resilience	% of subjects with redesigned tasks and inclusion of AI Declaration	Curricular reviews by Department	Semesterly

Note: Operational matrix proposed for the governance and follow-up of the pedagogical AI strategy

Conclusions

The present systematic literature review, elaborated under the PRISMA 2020 methodology over the five-year period 2020–2026, demonstrates that Artificial Intelligence has ceased to be a peripheral emerging technology in order to become a central infrastructure of higher education.

The fundamental findings confirm that:

1. Large Language Models and Generative AI are the dominant technologies (75.6 %), with near-universal student adoption (95 %), focused on personalised learning and academic support.
2. There exists a profound gap between automation and pedagogical collaboration, where 45.56 % of the studies reflect Substitution practices and fewer than 5 % achieve true Human–AI Synergy.
3. The positive impact of AI is not automatic; it requires explicit pedagogical mediation (PMAISE model), the development of teacher competences in the AI-TPACK framework and the redesign of assessments towards authentic and experiential models.
4. The ethical sustainability of the educational ecosystem requires surpassing punitive surveillance, implementing transparent, participatory governance policies adapted to the cultural realities of each institution.

In the final analysis, the future of higher education in the era of AI will not depend on the technical sophistication of the algorithms, but on the capacity of university institutions to situate human agency, critical thinking and pedagogical equity at the centre of technological innovation.

The five-phase operational strategy described in this article would provide higher education institutions with a pragmatic roadmap for transcending the prohibitive reaction and advancing towards a sustainable integration that prioritises critical judgement, equity of access and teacher professional development.

Future Research Agenda

The present systematic review identifies methodological gaps in the current literature that must be addressed by future research:

1. Longitudinal Designs: Surpass the overabundance of cross-sectional and short-duration exploratory studies in order to measure the impact of AI on knowledge retention and the development of professional competences in the long term.
2. Contextualised Studies in the Global South: Broaden the empirical evidence in universities of Latin America, Africa and Southern Asia, evaluating infrastructure barriers and the cultural pertinence of generative models.

3. Research on Human–AI Synergy: Develop and evaluate empirically platforms that foster the Synergy quadrant (Elsayed et al., 2026), objectively measuring learning gains when the teacher and AI collaborate dynamically.

Acknowledgements

The present work has been developed in the context of PI code 22F017, accredited by the General Secretariat of Science and Technique of the National University of the Northeast and PI No. 199, accredited by the Secretariat of Research, Science and Technique of the National University of Chaco Austral.

Conflict of Interest: The authors reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The authors did not obtain any funding for this research.

References:

1. Aldreabi, H., Dahdoul, N. K. S., Alhur, M., Alzboun, N., & Alsalhi, N. R. (2025). Determinants of student adoption of generative AI in higher education. Electronic Journal of e-Learning, 23(1), 15–33. https://doi.org/10.34190/ejel.23.1.3599
2. Allison, J., Hwang, G. J., Mayer, R. E., Pellas, N., Karnalim, O., de Freitas, S., … Sanusi, I. (2025). From generative AI to extended reality: Multidisciplinary perspectives on the challenges, opportunities, and future of educational computing. Journal of Educational Computing Research, 63(6), 1327–1363. https://doi.org/10.1177/07356331251359964
3. Amofa, B., Kamudyariwa, X., Fernandes, F., Osoba-jo, O., Jeremiah, F., & Oke, A. (2025). Navigating the complexity of generative artificial intelligence in higher education: A systematic literature review. Education Sciences, 15(7), 826. https://doi.org/10.3390/educsci15070826
4. Baig, M. I., & Yadegaridehkordi, E. (2024). ChatGPT in higher education: A systematic literature review and research challenges. International Journal of Educational Research, 127, Article 102411. https://doi.org/10.1016/j.ijer.2024.102411
5. Bond, M., Khosravi, H., de Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. International Journal

of Educational Technology in Higher Education, 21, 4. https://doi.org/10.1186/s41239-023-00436-z

6. Bretan, B. (2024). Teaching and learning legal English with AI: A case study on student engagement in an ESAP course. Studia Universitatis Babeș-Bolyai Philologia, 69, 153–168. https://doi.org/10.24193/subbphilo.2024.4.07
7. Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138, 107468. https://doi.org/10.1016/j.chb.2022.107468
8. Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20, 38. https://doi.org/10.1186/s41239-023-00408-3
9. Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20(1), 43. https://doi.org/10.1186/s41239-023-00411-8
10. Chan, C. K. Y., & Lee, K. K. (2023). The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial generation teachers? Smart Learning Environments, 10(1), 60. https://doi.org/10.1186/s40561-023-00269-3
11. Cheong, B. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. Frontiers in Human Dynamics, 6, Article 1421273. https://doi.org/10.3389/fhumd.2024.1421273
12. Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155
13. Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
14. Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, 22. https://doi.org/10.1186/s41239-023-00392-8
15. Dakakni, D., & Safa, N. (2023). Artificial intelligence in the L2 classroom: Implications and challenges on ethics and equity in higher education: A 21st century Pandora's box. Computers and Education: Artificial Intelligence, 5, Article 100179. https://doi.org/10.1016/j.caeai.2023.100179

16. Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. Computers & Education, 210, 104967. https://doi.org/10.1016/j.compedu.2023.104967
17. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008
18. Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. International Journal for Educational Integrity, 19(1), 25. https://doi.org/10.1007/s40979-023-00144-1
19. Elsayed, A., Malkawi, G., Wardat, Y., & Kamalov, F. (2026). Beyond automation: A systematic review of AI teaching methodologies and a framework for human–AI synergy in higher education. IEEE Access, 14, 45948–45963. https://doi.org/10.1109/ACCESS.2026.3675937
20. Evangelista, E. D., & Salman Bukhari, S. M. (2026). From ethical principles to executable governance: A policy-as-code framework for trustworthy AI in higher education. Computers and Education: Artificial Intelligence, 10, 100594. https://doi.org/10.1016/j.caeai.2026.100594
21. Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2023). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots' impact on university teachers' assessment practices. Assessment & Evaluation in Higher Education. https://doi.org/10.1080/02602938.2023.2241676
22. Floridi, L. (2019). Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 1(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y
23. Freeman, J. (2025). Student Generative AI Survey 2025 (HEPI Policy Note 61). Higher Education Policy Institute. https://www.hepi.ac.uk/reports/student-generative-ai-Survey-2025/
24. García-Carreño, I. V. (2025). Exploring risks and benefits in generative artificial intelligence through systematic review and bibliometric analysis. The European Educational Researcher, 8(3), 57–94. https://doi.org/10.31757/euer.834
25. Gouia-Zarrad, R., & Gunn, C. (2024). Enhancing students' learning experience in mathematics class through ChatGPT. International Electronic Journal of Mathematics Education, 19(3), em0781. https://doi.org/10.29333/iejme/14614
26. Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., et al. (2008). GRADE: An emerging consensus on

rating quality of evidence and strength of recommendations. BMJ, 336, 924. https://doi.org/10.1136/bmj.39489.470347.AD

27. He, W., Qiu, X., & Liao, J. (2024). Integration of traditional Chinese medicine constitution identification robot into nursing informatics curriculum: A case study at Xiangnan University. Frontiers in Artificial Intelligence and Applications, 387, 115–123. https://doi.org/10.3233/FAIA240249
28. Hofstede, G. (1980). Culture's consequences: International differences in work-related values. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000046536
29. Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information, 34(4), 285–291. https://doi.org/10.3233/EFI-180221
30. Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. British Journal of Educational Technology, 54(5), 1057–1076. https://doi.org/10.1111/bjet.13325
31. Kendeou, P., Greene, J. A., Nixon, N., Choi, J. Y., Fulsher, A., Choi, J., Aguilar, S., Allen, L., Azevedo, R., Bernacki, M., Butler, A., Carpenter, S., Crompton, H., Doroudi, S., Fiorella, L., Fyfe, E., Graesser, A., Hutt, S., Karumbaiah, S., Kizilcec, R., Lee, V., McNamara, D., Miele, D. B., Mills, C., Muis, K. R., Roschelle, J., Roscoe, R. D., Stewart, A., Walker, E., Wang, X. C., Witherby, A., Xu, Y., & Zhai, X. (2026). Ways generative AI can support and threaten learning in K-20 US education (Expert Consensus Report). https://cehd.umn.edu/AI-expert-consensus-report-2026
32. Koehler, M. J., Mishra, P., & Cain, W. (2013). What is Technological Pedagogical Content Knowledge (TPACK)? Journal of Education, 193(3), 13–19. https://doi.org/10.1177/002205741319300303
33. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310
34. Lasfeto, D. B., & Ulfa, S. (2023). Modeling of online learning strategies based on fuzzy expert systems and self-directed learning readiness: The effect on learning outcomes. Journal of Educational Computing Research, 60(8), 2081–2104. https://doi.org/10.1177/07356331221094249
35. Li, S., & Gu, X. (2023). A risk framework for human-centered artificial intelligence in education: Based on literature review and

Delphi-AHP method. Educational Technology & Society, 26(1), 187–202. https://doi.org/10.30191/ETS.202301_26(1).0014

36. Linden, K., Pemberton, L., & Webster, L. (2019). Evaluating the bones of adaptive learning: Do the initial promises really increase student engagement and flexible learning within first year anatomy subjects? In J. Domenech, P. Merello, E. De La Poza, D. Blazquez, & R. Pena Ortiz (Eds.), 5th International Conference on Higher Education Advances (HEAd'19) (pp. 331–339). Charles Sturt University. https://dialnet.unirioja.es/servlet/articulo?codigo=9576874
37. Long, D., Wang, S., Md Rashid, S., & Lu, X.-T. (2026). Artificial intelligence in higher education: A systematic review of its impact on student engagement and the mediating role of teaching methods. Frontiers in Education, 10, 1648661. https://doi.org/10.3389/feduc.2025.1648661
38. Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. Assessment & Evaluation in Higher Education, 49(5), 651–664. https://doi.org/10.1080/02602938.2024.2309963
39. Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x
40. Moorhouse, B. L. (2024). Beginning and first-year language teachers' readiness for the generative AI age. Computers and Education: Artificial Intelligence, 6, 100201. https://doi.org/10.1016/j.caeai.2024.100201
41. Newton, P., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education: A pragmatic scoping review. Assessment & Evaluation in Higher Education, 49(6), 781–798. https://doi.org/10.1080/02602938.2023.2299059
42. Ning, Y., Zhang, C., Xu, B., Zhou, Y., & Wijaya, T. T. (2024). Teachers' AI-TPACK: Exploring the relationship between knowledge elements. Sustainability, 16(3), 978. https://doi.org/10.3390/su16030978
43. Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. International Journal of Educational Technology in Higher Education, 20, 4. https://doi.org/10.1186/s41239-022-00372-4

44. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
45. Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching and Learning Practice, 20(2). https://doi.org/10.53761/1.20.02.07
46. Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning, 12(1), 22. https://doi.org/10.1186/s41039-017-0062-8
47. Prather, J., et al. (2023). The robots are here: Navigating the generative AI revolution in computing education. Proceedings of the ITiCSE Working Group Reports on Innovation and Technology in Computer Science Education, 108–153. https://doi.org/10.1145/3623762.3633499
48. Ren, X., & Wu, M. (2025). Examining teaching competencies and challenges while integrating artificial intelligence in higher education. TechTrends, 69(3), 519–538. https://doi.org/10.1007/s11528-025-01055-3
49. Rivas Vera, T. M., Zambrano Zambrano, W. V., & Ramos Lazo, M. J. (2026). Impacto de la inteligencia artificial aplicada en los procesos de enseñanza y aprendizaje en educación superior: Una revisión sistemática. RECIMUNDO, 10(4). https://doi.org/10.26820/recimundo/10.(4).nov.2026.331-355
50. Rodríguez-Abitia, G., Martínez-Pérez, S., Ramirez-Montoya, M. S., & Lopez-Caudana, E. (2020). Digital gap in universities and challenges for quality education: A diagnostic study in Mexico and Spain. Sustainability, 12(21), 9069. https://doi.org/10.3390/su12219069
51. Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68
52. Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Using generative artificial intelligence tools to explain and enhance experiential learning for authentic assessment. Education Sciences, 14, 83. https://doi.org/10.3390/educsci14010083

53. Stephenson, R., & Armstrong, C. (2026). Student Generative AI Survey 2026. Higher Education Policy Institute. https://www.hepi.ac.uk/reports/student-generative-ai-survey-2026/
54. Sun, D., Ba, S., Cha, Y., Yu, J., Chiang, F.-K., Dai, H. M., & Lim, C.-P. (2026). Empowering university teachers in higher education: A generative AI-responsive competency framework. Computers and Education: Artificial Intelligence, 10, 100542. https://doi.org/10.1016/j.caeai.2026.100542
55. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://doi.org/10.48550/arXiv.1706.03762
57. Xi, L., Zhang, Y., & Wang, Q. (2026). Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. Computers & Education, 241, Article 105494. https://doi.org/10.1016/j.compedu.2025.105494
58. Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. Languages, 8(3), 212. https://doi.org/10.3390/languages8030212
59. Yan, S., Yan, Y., Wang, S., Li, Y., & Ai, Z. (2026). Learning with AI, not from AI: A quasi-experimental comparison of AI-assisted learning with versus without an interactive AI-generated error-correction phase on knowledge application and metacognition. International Journal of Educational Research, 139, 103047. https://doi.org/10.1016/j.ijer.2026.103047
60. Yusuf, A., Pervin, N., & Román-González, M. (2024). Generative AI and the future of higher education: A threat to academic integrity or reformation? Evidence from multicultural perspectives. International Journal of Educational Technology in Higher Education, 21, 21. https://doi.org/10.1186/s41239-024-00453-6