



Generative AI in Higher Education: Evidence Synthesis of Applied Learning Outcomes

Muhammad Nazirul Mubin Bin Abdul Moen^{1*}, Alfa Rohmatin², Mia Siti Nurazizah³, Mehwish Kaleem⁴

¹ International Islamic University Malaysia, Malaysia

^{2,3} STIT Al Azami Cianjur, Indonesia

⁴ Muslim Youth (MY) University, Pakistan

Email: nazirulmubin.m@live.iium.edu.my¹

Article Info :

Received:

18-09-2026

Revised:

14-10-2026

Accepted:

20-10-2026

Abstract

Generative artificial intelligence has spread through higher education faster than institutions have produced stable evidence on learning effects. This study synthesized recent empirical reviews, meta-analyses, and applied higher-education research to identify where generative artificial intelligence improved learning and where outcomes remained contingent on instructional design. A structured evidence-synthesis method compared five quantitative or systematic anchor reviews with complementary studies on assessment, feedback, academic integrity, and learner regulation. The anchor literature ranged from 34 to 103 included studies, while a 2026 three-level meta-analysis reported an overall learning effect of $g = 0.499$ across 36 studies, 132 effect sizes, and 7,229 participants. Larger effects were reported for collaborative learning and cognitive or creative outcomes, while other reviews showed positive achievement and motivation effects but inconsistent self-efficacy findings. The synthesis concluded that generative artificial intelligence produced stronger outcomes when embedded in guided, task-specific pedagogy rather than used as an unrestricted answer generator. Assessment redesign, verification routines, and learner metacognition remained necessary implementation conditions.

Keywords: assessment, generative artificial intelligence, higher education, learning outcomes, evidence synthesis.



©2026 Authors.. This work is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License.

(<https://creativecommons.org/licenses/by-nc/4.0/>)

INTRODUCTION

Generative artificial intelligence entered higher education as a general-purpose writing, coding, explanation, and feedback technology rather than as a conventional educational application designed around a fixed curriculum. Large language models can produce fluent responses, adapt explanations, generate examples, and support iterative dialogue, creating immediate opportunities for tutoring and feedback but also risks related to hallucination, overreliance, opacity, bias, and academic integrity (Kasneci et al., 2023; Tlili et al., 2023). The speed of adoption has complicated institutional decision-making because policy questions emerged before universities had accumulated longitudinal evidence on learning gains. OpenAI (2023) documented the broad capabilities and limitations of GPT-4, while UNESCO (2023) emphasized human oversight, age-appropriate use, privacy, and pedagogical purpose. The applied research problem is therefore no longer whether students can access generative systems, but under which instructional arrangements their use produces measurable educational benefit rather than merely faster task completion.

Early higher-education literature was dominated by conceptual discussion and educator perception studies. Lo (2023) mapped rapid growth in educational uses and concerns, while Dwivedi et al. (2023) assembled multidisciplinary perspectives on implications for research, practice, and policy. Lim et al. (2023) framed generative artificial intelligence as both an opportunity and a challenge for management education, especially where assessment tasks can be completed through generic prompting. Cotton et al. (2024) focused on academic integrity and the need to redesign assessment practices rather than relying on detection alone. These contributions established a useful problem map, yet institutions still require quantitative evidence about achievement, motivation, higher-order thinking, collaboration, and self-regulation. A synthesis centered on applied learning outcomes is necessary because technical capability, perceived usefulness, and actual learning effectiveness are analytically distinct.

The empirical literature has expanded sufficiently to permit more disciplined comparison. Weng et al. (2024) synthesized 34 studies on generative artificial intelligence in education and documented growing attention to learning support, feedback, and instructional design. Baig and Yadegaridehkordi (2024) reviewed 57 studies and highlighted both educational affordances and implementation challenges. Hawi and Samaha (2026) later examined 103 empirical studies, indicating that the evidence base had moved beyond a small collection of exploratory reports. Meta-analytic work has also begun to estimate outcome magnitude. Fan et al. (2026) analyzed 36 studies, 132 effect sizes, and 7,229 participants using a three-level model, reporting an overall effect of $g = 0.499$ and stronger effects for several specific learning configurations. These syntheses enable a comparative assessment of where evidence converges and where it remains heterogeneous.

Learning outcomes cannot be interpreted without considering the design of student–AI interaction. Kohnke et al. (2023) show how generative systems can support language learning through dialogic practice and resource generation, while Jeon and Lee (2023) emphasize the changing roles of teachers and chatbots in language education. Rahman and Watanobe (2023) discuss the educational possibilities of ChatGPT alongside risks of inaccurate information and dependency. Moorhouse et al. (2023) likewise demonstrate that language-learning use requires teacher mediation to align generated material with learning objectives. The common mechanism across these studies is guided augmentation: students obtain explanations, examples, critique, or feedback while retaining responsibility for reasoning and verification. This differs substantially from replacement use, where the model performs the core intellectual task and the learner simply submits or lightly edits the output.

Assessment has become the principal institutional pressure point because generative systems weaken the assumption that submitted text directly represents unaided student performance. Rudolph et al. (2023) argue that traditional take-home assessment practices require reconsideration, while Chan and Hu (2023) document student perceptions that combine strong perceived usefulness with concerns about accuracy and ethics. Farrokhnia et al. (2024) use a SWOT-based analysis to show that educational value depends on policy, literacy, and task design rather than on the tool alone. Grassini (2023) similarly identifies both opportunities for personalized support and risks to independent thinking. Policy guidance from the European Commission (2022) and OECD (2023) reinforces a human-centered approach that combines transparency, teacher judgment, data protection, and explicit learning goals. These sources suggest that the relevant unit of evaluation is the pedagogical arrangement, not the model in isolation.

This study addressed a gap between rapidly accumulating evidence reviews and practical higher-education implementation. It synthesized quantitative findings from recent systematic reviews and meta-analyses, then interpreted those findings using complementary research on assessment, feedback, integrity, and student regulation. The objectives were to compare the size and direction of reported learning effects, identify instructional conditions associated with stronger outcomes, and develop an applied implementation framework for university courses. The review was deliberately structured rather than claimed as an exhaustive PRISMA systematic review because the analytical purpose was triangulation across published syntheses and high-relevance empirical literature, not *de novo* screening of every primary study. This approach is consistent with the need identified by Zawacki-Richter et al. (2019) to connect artificial-intelligence research with pedagogical practice and with Sallam (2023) to appraise benefits and risks using transparent evidence rather than technological enthusiasm.

RESEARCH METHODS

A structured evidence-synthesis design was used. Five anchor sources were selected because they reported either systematic-review study counts, quantitative meta-analytic outcomes, or both: Weng et al. (2024), Baig and Yadegaridehkordi (2024), Deng et al. (2025), Liu et al. (2025), Fan et al. (2026), and the broader empirical mapping by Hawi and Samaha (2026). The anchor set was complemented by peer-reviewed studies and reviews addressing higher-education use cases, assessment, academic integrity, language learning, and learner perceptions. Eligible evidence had to discuss generative artificial intelligence or large language models in educational settings and provide an identifiable implication for learning, teaching, assessment, or implementation. Sources focused solely on technical model architecture without an educational outcome were excluded from the analytical narrative.

Data extraction recorded publication year, review type, number of included studies when reported, participant counts where available, outcome domains, effect estimates, and implementation moderators. Quantitative effects were not pooled again because the underlying reviews used different eligibility

rules, effect metrics, and study populations. Instead, effect sizes were preserved in the form reported by the original synthesis and compared descriptively. Fan et al. (2026) provided the most detailed cross-study quantitative benchmark, including an overall g of 0.499, $g = 0.504$ for higher-order thinking, $g = 0.669$ for cognitive or creative outcomes, $g = 1.026$ for collaborative learning, and $g = 0.633$ for blended learning. Liu et al. (2025) reported effects of 0.857 for learning achievement and 0.803 for motivation, while Deng et al. (2025) reported positive effects across several outcome families but no statistically significant improvement in self-efficacy.

The interpretive coding scheme separated four mechanisms: cognitive support, feedback and iteration, motivational or affective support, and collaborative or dialogic support. A second layer coded risks that could weaken observed learning, including answer substitution, factual unreliability, shallow prompting, unverified citations, assessment misalignment, and reduced metacognitive monitoring. The scheme was informed by empirical and conceptual work on generative-AI pedagogy (Bhullar et al., 2024; Albadarin et al., 2024; Chiu et al., 2024) and by broader reviews of artificial intelligence in higher education (Bond et al., 2024; Zawacki-Richter et al., 2019). The analysis therefore treated effect estimates as outcomes produced within instructional systems rather than as stable properties of generative artificial intelligence itself.

Robustness was strengthened through source triangulation and explicit separation of findings from interpretation. Results reported by the anchor syntheses were reproduced only at the level necessary for comparison, and contradictory or nonsignificant findings were retained. Background literature was used to explain plausible pedagogical mechanisms rather than to inflate the quantitative evidence base. Because the field changes quickly, publication date was treated as an important contextual variable, and the synthesis included evidence available through 2026. The method does not estimate publication bias across the entire primary literature and cannot resolve dependence where the same primary studies appear in multiple reviews. Its purpose is an applied evidence map suitable for institutional design, consistent with the cautious implementation principles advanced by UNESCO (2023), OECD (2023), and the European Commission (2022).

RESULTS AND DISCUSSION

Magnitude and distribution of reported learning effects

The first result is that the evidence base has grown from exploratory commentary into a sizable body of empirical synthesis. The anchor reviews ranged from 34 included studies in Weng et al. (2024) to 103 empirical studies in Hawi and Samaha (2026), with Baig and Yadegaridehkordi (2024) reviewing 57 studies and Liu et al. (2025) synthesizing 49 articles. Study counts are not directly comparable because eligibility criteria differ, but the sequence shows that the literature has accumulated rapidly across only a few academic years. Table 1 summarizes the scope and principal contribution of the anchor reviews. The expansion matters because early claims about generative artificial intelligence were often based on isolated classroom demonstrations; later syntheses can distinguish recurring outcomes from tool-specific anecdotes. Bhullar et al. (2024) and Albadarin et al. (2024) similarly report a shift toward empirical examination of educational practice, although methodological heterogeneity remains substantial.

Table 1. Selected Evidence Syntheses on Generative AI in Education

Study	Evidence base	Principal quantitative or analytical contribution
Weng et al. (2024)	34 studies	Systematic mapping of pedagogical applications
Fan et al. (2026)	36 studies; 132 effects; $n=7,229$	Three-level meta-analysis; overall $g=0.499$
Liu et al. (2025)	49 articles	Achievement=0.857; motivation=0.803

Study	Evidence base	Principal quantitative or analytical contribution
Baig & Yadegaridehkordi (2024)	57 studies	Systematic review of opportunities and challenges
Hawi & Samaha (2026)	103 empirical studies	Broad empirical synthesis across educational uses

Source: Compiled from the cited reviews and meta-analyses.

Fan et al. (2026) provide the clearest quantitative benchmark in the synthesized evidence. Their three-level meta-analysis covered 36 studies, 132 effect sizes, and 7,229 participants and produced an overall effect of $g = 0.499$. An effect of this magnitude is educationally meaningful but does not support the claim that generative artificial intelligence produces uniformly large gains. The three-level approach is especially relevant because multiple effects can be nested within the same study, a dependence structure that simpler meta-analytic summaries may not handle adequately. The result suggests that, on average, generative-AI-supported interventions improved measured learning relative to comparison conditions, but the variation across outcomes and instructional designs is at least as important as the pooled mean. Figure 1 positions this overall estimate beside selected domain effects reported in the same synthesis.

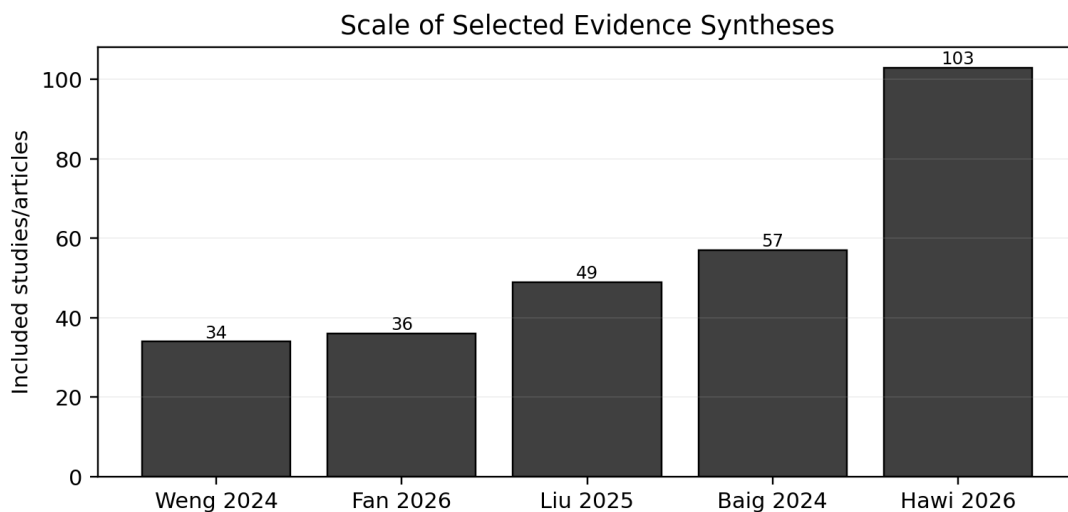


Figure 1. Scale of Selected Generative-AI Evidence Syntheses

Outcome-specific estimates indicate stronger effects where the technology supports complex interaction rather than passive answer retrieval. Fan et al. (2026) reported $g = 0.504$ for higher-order thinking and $g = 0.669$ for cognitive or creative outcomes. These findings are compatible with instructional uses that ask students to critique generated explanations, compare alternative solutions, iterate drafts, or use the model as a dialogic partner. Kasneci et al. (2023) anticipated this mechanism by emphasizing personalized explanation and feedback as promising educational affordances, while Kohnke et al. (2023) described dialogic language-learning applications that require learner participation. The result does not mean that any prompting activity improves critical thinking. It indicates that interventions explicitly designed around reasoning can achieve measurable gains when the learner still performs evaluative work.

The largest selected effect in Fan et al. (2026) concerned collaborative learning, with $g = 1.026$. This estimate should be interpreted cautiously because domain-level effects can be based on smaller subsets and may be sensitive to study design, but it suggests that generative systems can function as scaffolds in group inquiry, role-play, brainstorming, and peer-feedback arrangements. Jeon and Lee (2023) show that productive educational use often involves redistribution of roles among teacher,

learner, and chatbot rather than simple automation. A model can supply prompts, perspectives, counterarguments, or examples that keep group interaction moving, while human participants retain responsibility for evaluating relevance and quality. This pattern also helps explain why pedagogically orchestrated use may outperform unrestricted individual use: social accountability and teacher framing create external checks on generated content.

Blended learning also produced a comparatively strong effect in Fan et al. (2026), $g = 0.633$. Blended designs can combine model-assisted practice with teacher explanation, peer interaction, formal resources, and assessment, reducing the risk that the model becomes the sole epistemic authority. The structure is consistent with OECD (2023) guidance that digital tools should be integrated into teaching systems rather than treated as substitutes for instructional expertise. It also aligns with the human-centered emphasis in the European Commission (2022), where educators retain responsibility for learning objectives, fairness, and professional judgment. The evidence therefore supports a conditional proposition: generative artificial intelligence appears more educationally productive when embedded within a broader sequence of activities that includes independent reasoning and external verification.

Liu et al. (2025) reported large positive effects for learning achievement (0.857) and motivation (0.803) across 49 articles. These estimates are larger than the overall value reported by Fan et al. (2026), illustrating why effect sizes from different syntheses should not be treated as if they came from one common statistical model. Differences in populations, intervention duration, control conditions, outcome instruments, and effect metrics can produce substantial variation. Even so, the direction of results supports the view that generative tools can improve both performance and engagement under appropriate conditions. Chan and Hu (2023) found that students commonly perceive benefits related to efficiency and learning support, which provides a plausible motivational pathway. The practical implication is to preserve engagement benefits while designing tasks that still demand explanation, evidence, and original judgment.

Deng et al. (2025) add an important qualification by reporting positive effects on academic performance, affective-motivational outcomes, and higher-order thinking, but no significant effect on self-efficacy. This divergence is theoretically coherent. A student may complete tasks more successfully with immediate model assistance without becoming more confident in independent performance, particularly when the student attributes success to the tool rather than to personal mastery. Grassini (2023) and Farrokhnia et al. (2024) both identify overreliance as a risk that can accompany convenience. A mature evaluation strategy should therefore measure transfer: whether students can perform related tasks without assistance after using generative tools during learning. Achievement during assisted practice is valuable, but durable competence requires evidence that skills persist when scaffolding is reduced.

Figure 2 compares selected effect magnitudes while preserving their source-specific status. The visual pattern makes two points visible: reported effects are generally positive, and heterogeneity is substantial. The largest values occur in collaborative learning and in the achievement and motivation estimates from Liu et al. (2025), whereas the pooled overall effect from Fan et al. (2026) is moderate. Such heterogeneity is expected because 'generative AI use' is not a single treatment. It includes tutoring, question generation, automated feedback, ideation, code assistance, translation, simulation, and writing support, each interacting with course level, discipline, assessment structure, and student expertise. Baig and Yadegaridehkordi (2024) emphasize this diversity, making it inappropriate to collapse all applications into a simple claim that the technology is effective or ineffective.

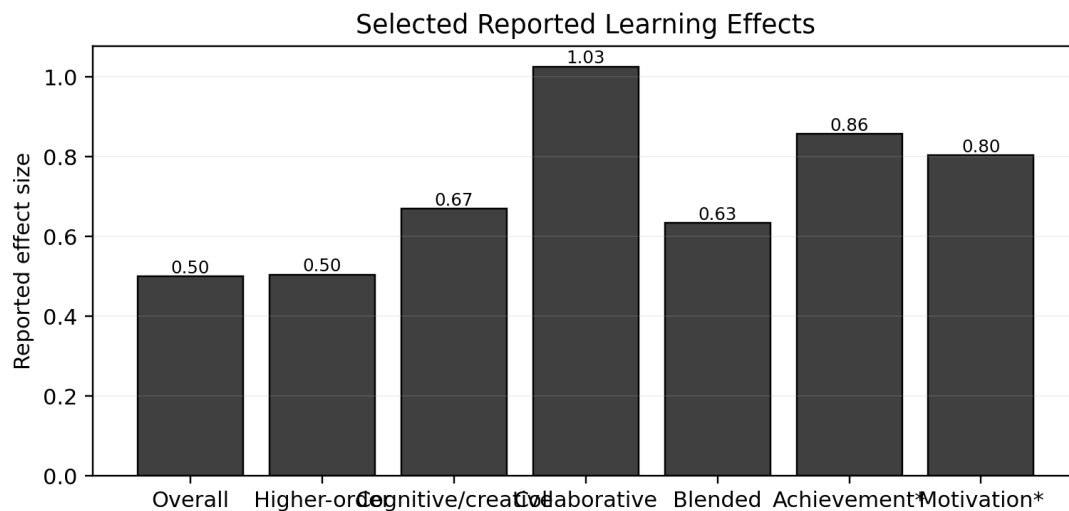


Figure 2. Selected Reported Learning Effects

The evidence base also shows a timing problem that matters for interpretation. Many studies were conducted during rapid model turnover, so an intervention evaluated with one generation of a system may not represent the capabilities, interface constraints, or safeguards available a year later. Lo (2023) and Tlili et al. (2023) documented the speed with which early educational uses expanded, while Chan and Hu (2023) showed that student perceptions already combined practical enthusiasm with concerns about accuracy, ethics, and dependence. This temporal instability strengthens the case for reporting the model, access mode, instructional prompt structure, duration, and assessment conditions in empirical studies instead of treating generative artificial intelligence as a stable intervention category. It also explains part of the heterogeneity across syntheses: effect estimates aggregate studies conducted under materially different technological and pedagogical conditions. For applied decision-making, effect direction is informative, but implementation details remain essential for judging whether a reported result is transferable to a particular course or institution.

The quantitative evidence therefore supports a constrained conclusion rather than a universal one. Generative-AI-supported learning can improve achievement, motivation, higher-order thinking, creativity, and collaboration, but effect magnitude depends on task design and outcome definition. The strongest educational interpretation is that the technology acts as a variable scaffold whose value rises when students are required to interact, verify, revise, and explain. This conclusion is consistent with the broader empirical mapping by Hawi and Samaha (2026), which documents extensive but heterogeneous implementation, and with Weng et al. (2024), which identifies design and pedagogical integration as recurring themes. The next subsection examines the conditions that convert technical capability into learning and the conditions that can reverse that benefit.

Instructional conditions, assessment design, and responsible implementation

The synthesis identified structured prompting and explicit learning goals as the first implementation condition. When students are asked to obtain an answer with no requirement to expose reasoning, compare evidence, or revise a product, the model can replace the very cognitive process an assignment is intended to develop. By contrast, prompts that require students to generate alternatives, identify assumptions, critique model output, and justify final choices create observable reasoning traces. Rahman and Watanobe (2023) note the potential of ChatGPT for explanation and practice but also emphasize reliability concerns, making verification an inseparable component of learning design. Chiu et al. (2024) similarly argue that artificial-intelligence-supported education requires a clear alignment between technological functions and pedagogical goals. Table 2 translates this principle into task-level design options.

Table 2. Task Design Patterns for Learning-Oriented Generative AI Use

Design feature	Lower-learning-value pattern	Learning-oriented pattern
Prompting	Request final answer	Compare alternatives and justify selection
Feedback	Accept model revision	Evaluate feedback against rubric and revise
Evidence	Use generated citations	Retrieve and verify primary sources
Reasoning	Submit generated explanation	Annotate assumptions, errors, and reasoning steps
Transfer	AI available for every task	Alternate assisted and unassisted performance

Source: Authors' synthesis of the cited pedagogical evidence.

Feedback is the second mechanism with strong applied potential because generative systems can respond immediately and repeatedly. Students can request explanation of errors, alternative examples, or critique of a draft before formal submission, reducing the latency that limits human feedback in large classes. Moorhouse et al. (2023) show how teachers can integrate generative tools into language learning while still curating and contextualizing output. Kohnke et al. (2023) likewise describe conversational applications that support practice and feedback. The educational benefit depends on whether students use feedback to revise and whether teachers provide criteria for judging model suggestions. A high-volume feedback channel that students accept uncritically can spread misconceptions quickly; a guided feedback loop can instead increase opportunities for deliberate practice.

Assessment redesign is the third condition and the most urgent institutional challenge. Cotton et al. (2024) argue that generative artificial intelligence complicates conventional assumptions about authorship and academic integrity, while Rudolph et al. (2023) contend that universities should reconsider assessment formats that can be completed through generic text generation. A defensible response is to increase process evidence: staged drafts, oral explanation, annotated source use, in-class components, data-specific analysis, and reflective accounts of how tools were used. These approaches assess both the product and the learner's command of the underlying reasoning. They also reduce the incentive to treat detection software as the primary integrity mechanism, an approach that can produce false accusations and does little to improve learning design.

Verification literacy emerged as a fourth condition. Large language models can produce fluent but inaccurate claims, fabricated references, or overconfident interpretations, which means students need explicit routines for checking factual statements and source provenance. Sallam (2023) documented this concern across early ChatGPT literature, and OpenAI (2023) acknowledged limitations involving hallucination and reliability. Verification can be operationalized through source retrieval, cross-checking with primary literature, calculation replication, and disclosure of uncertain claims. The educational opportunity is significant: rather than treating hallucination only as a defect, instructors can use it to teach evidence evaluation and epistemic vigilance. Such design turns model fallibility into a structured critical-thinking task while preserving responsibility with the student.

Learner regulation is the fifth condition. Generative systems reduce effort for drafting, searching, and explanation, but reduced effort is not always educationally desirable because productive struggle can contribute to learning. If students outsource planning, monitoring, and evaluation, short-term performance may rise while independent competence stagnates. Deng et al. (2025) provide a relevant signal through the absence of a significant self-efficacy effect despite positive outcomes elsewhere. Albadarin et al. (2024) also highlight self-regulation and learner agency as important dimensions of generative-AI-supported learning. Course designs can address this by alternating assisted and unassisted

tasks, requiring students to predict answers before consulting the model, and assessing transfer under conditions where AI support is limited or unavailable.

Institutional policy should distinguish permitted use by learning purpose rather than relying on a binary allowed/prohibited rule. A brainstorming exercise, formative tutoring dialogue, coding-debugging task, and final individual examination carry different risks and learning objectives. UNESCO (2023) recommends governance that considers human agency, privacy, transparency, and educational appropriateness, while OECD (2023) emphasizes system-level capacity for responsible digital transformation. A purpose-based policy can state what assistance is allowed, what must be disclosed, which data may not be uploaded, and which parts of the work must remain demonstrably independent. This level of specificity also supports fairness because students receive common expectations instead of relying on informal norms that vary across instructors.

Faculty capability is equally important because assessment redesign and verification cannot be delegated to students alone. Educators need enough model literacy to design assignments that expose reasoning, anticipate likely AI-generated shortcuts, and evaluate disclosed tool use. Bond et al. (2024) show that artificial-intelligence research in higher education has often developed separately from teacher-centered pedagogical expertise, echoing the earlier finding of Zawacki-Richter et al. (2019) that educators were underrepresented in AI research. Closing that gap requires professional development focused on concrete course design rather than generic demonstrations of model capabilities. The objective is not to turn every lecturer into a machine-learning specialist; it is to ensure that teaching decisions remain grounded in disciplinary standards and credible evidence about learning.

Table 3 consolidates the evidence into an implementation matrix. High-value uses are those in which AI increases opportunities for practice, dialogue, feedback, or comparison while the student remains accountable for verification and final reasoning. Higher-risk uses are those in which the model generates the assessed intellectual product with little process evidence or where sensitive institutional and personal data are entered into systems without appropriate safeguards. Chan and Hu (2023) show that students themselves recognize both benefits and concerns, suggesting that transparent course-level rules can be aligned with learner expectations. Lim et al. (2023) likewise argue that universities need to balance innovation with academic standards. The matrix is intended as a design aid rather than a universal rule because disciplines differ in what constitutes authentic performance.

Table 3. Applied Implementation Matrix for Higher Education

Use case	Potential learning value	Primary control
Formative tutoring	High when explanations are interrogated	Verification prompts and teacher-aligned criteria
Draft feedback	High when followed by revision	Version history and reflective commentary
Brainstorming	Moderate to high	Independent selection and justification
Summative essay generation	Low if product replaces reasoning	Process evidence, oral validation, disclosure
Coding/debugging	High for iterative practice	Explain code, tests, and error diagnosis
Sensitive-data tasks	Context dependent	Privacy restrictions and approved systems

Source: Authors' synthesis; categories are implementation guidance, not a performance ranking.

Responsible implementation also requires a feedback cycle between course-level evidence and institutional governance. Universities can require disclosure of permitted tool use, sample student work for verification, monitor assessment anomalies, and review whether learning outcomes change after major model or policy updates. UNESCO (2023) frames human agency, transparency, privacy,

and accountability as core governance principles, while the European Commission (2022) places similar emphasis on educator judgment and ethical use of artificial intelligence in education. Lim et al. (2023) argue that higher education institutions need policies capable of adapting as generative systems evolve rather than relying on static prohibition or unrestricted adoption. A recurring review mechanism can connect those principles to measurable indicators such as revision quality, source-verification accuracy, unauthorized assistance cases, student confidence in independent work, and staff workload. This arrangement keeps governance tied to educational performance and allows institutions to revise rules when evidence shows that a particular use either strengthens or weakens the intended learning process.

The synthesis has limitations. Review overlap means that some primary studies may contribute to more than one published synthesis, intervention durations are often short, and rapidly changing model capabilities can make evidence temporally fragile. Effect sizes reported by different meta-analyses cannot be directly pooled without access to common primary-study data and harmonized coding. Longitudinal evidence on retention, transfer, and independent competence remains thinner than evidence on immediate task performance. Despite these constraints, the convergence across Fan et al. (2026), Liu et al. (2025), Deng et al. (2025), Baig and Yadegaridehkordi (2024), and Hawi and Samaha (2026) is sufficient to reject two simplistic positions: that generative artificial intelligence automatically improves learning, or that its educational use is inherently detrimental. The applied question is how institutions structure the interaction.

CONCLUSION

Recent evidence supports a conditional positive association between generative-artificial-intelligence-supported instruction and several higher-education learning outcomes. The strongest quantitative anchor in this synthesis reported an overall effect of $g = 0.499$ across 36 studies, 132 effects, and 7,229 participants, with larger estimates for collaborative, cognitive or creative, and blended-learning configurations. Other syntheses reported substantial gains in achievement and motivation while also showing that self-efficacy does not necessarily improve alongside assisted performance. These results make instructional design the central explanatory variable rather than technological access alone.

Universities can use generative systems most defensibly when tasks require visible reasoning, iterative feedback, source verification, and transfer to independent performance. Assessment redesign should combine process evidence, oral or in-class validation, disclosure rules, and discipline-specific standards instead of depending primarily on AI detection. Future research should prioritize longer interventions, common outcome measures, unassisted transfer tests, and transparent reporting of model version, prompting structure, instructor involvement, and student prior knowledge.

REFERENCES

- Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, 3, 60. <https://doi.org/10.1007/s44217-024-00138-2>
- Baig, M. I., & Yadegaridehkordi, E. (2024). Generative AI in education: A systematic review of opportunities, challenges, and research directions. *International Journal of Educational Research*, 127, 102411. <https://doi.org/10.1016/j.ijer.2024.102411>
- Bhullar, P. S., Joshi, M., & Chugh, R. (2024). ChatGPT in higher education: A synthesis of the literature and a future research agenda. *Education and Information Technologies*, 29, 21501–21522. <https://doi.org/10.1007/s10639-024-12723-x>
- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., et al. (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>
- 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, 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2024). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in

- education. Computers and Education: Artificial Intelligence, 4, 100118. <https://doi.org/10.1016/j.caeai.2022.100118>
- Cotton, D. R. E., 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>
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224. <https://doi.org/10.1016/j.compedu.2024.105224>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- European Commission. (2022). Ethical guidelines on the use of artificial intelligence and data in teaching and learning for educators. Publications Office of the European Union.
- Fan, Y., Ke, Q., Chen, L., & Lv, C. (2026). The effects of generative artificial intelligence on learning outcomes: A three-level meta-analysis. *Frontiers in Psychology*, 17, 1758670. <https://doi.org/10.3389/fpsyg.2026.1758670>
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 61(3), 460–474. <https://doi.org/10.1080/14703297.2023.2195846>
- Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, 13(7), 692. <https://doi.org/10.3390/educsci13070692>
- Hawi, N. S., & Samaha, M. (2026). Generative artificial intelligence in education: A systematic review of empirical evidence. *International Journal of Educational Research Open*, 11, 100584. <https://doi.org/10.1016/j.ijedro.2026.100584>
- Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28, 15873–15892.
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550. <https://doi.org/10.1177/00336882231162868>
- Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2), 100790. <https://doi.org/10.1016/j.ijme.2023.100790>
- Liu, Y., Chen, Y., & Li, H. (2025). Effects of generative artificial intelligence on student learning: A meta-analytic review. *Journal of Educational Computing Research*. <https://doi.org/10.1177/07356331251329185>
- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, 5, 100151. <https://doi.org/10.1016/j.caeo.2023.100151>
- OECD. (2023). OECD digital education outlook 2023: Towards an effective digital education ecosystem. OECD Publishing. <https://doi.org/10.1787/c74f03de-en>
- OpenAI. (2023). GPT-4 technical report. arXiv. <https://doi.org/10.48550/arXiv.2303.08774>
- Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9), 5783. <https://doi.org/10.3390/app13095783>
- Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching*, 6(1), 342–363. <https://doi.org/10.37074/jalt.2023.6.1.9>

- Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, 15. <https://doi.org/10.1186/s40561-023-00237-x>
- UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.
- Weng, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Generative AI in education: A systematic review of pedagogical applications and implications. *Australasian Journal of Educational Technology*, 40(6), 37–55. <https://doi.org/10.14742/ajet.9540>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>