The Human Moat: A Conceptual Framework for Preserving Distinctively Human Capabilities in the Generative AI Era, with Implications for Malaysian Higher Education

Mahazir Ismail[1], Mohd Firdaus Ruslan[1*], Kardina Kamaruddin[1]

Faculty of Business and Management, Universiti Teknologi MARA, Kedah, Malaysia

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.101000230

Received: 17 October 2026; Accepted: 22 October 2026; Published: 01 November 2026

ABSTRACT

Generative artificial intelligence (GenAI) can now carry out many of the cognitive tasks through which universities have long developed their students. The research that has followed is substantial but scattered, with separate strands on cognitive offloading, critical thinking, metacognition, human agency and AI literacy, and it has been far more concerned with documenting erosion than with specifying what ought to be protected. This conceptual paper proposes the Human Moat, defined as the set of distinctively human capabilities that learners keep as their own rather than delegate to GenAI, and whose value increases as AI takes over routine cognitive work. The framework rests on task-based theories of automation and on cognitive offloading theory, and it is grounded in 173 unique Web of Science and Scopus records (2024–2027) retrieved on 15 October 2026 and screened for recurring capability themes. Five dimensions of the moat are identified: critical and evaluative judgement, epistemic vigilance, metacognitive self-regulation, agency and ethical accountability, and creative generation. AI literacy is treated as an enabling condition, while pedagogical design is proposed as the switch that directs GenAI use towards either a moat-building (augmentative) or a moat-eroding (substitutive) pathway. The paper distinguishes the Human Moat from neighbouring constructs and states seven propositions for empirical testing. Malaysian higher education serves as an illustrative case, since national guidelines already caution that unregulated GenAI use may weaken critical thinking, yet only one record in the corpus had a Malaysia-affiliated author. Implications are drawn for curriculum, assessment, faculty development and institutional policy, together with an agenda for measuring the moat. To support application, the paper also sets out its framework-development procedure and selection criteria, proposes measurable indicators for each dimension, maps the dimensions to the Malaysian Qualifications Framework, compares the framework with international guidance, discusses institutional and cultural challenges and outlines a staged plan for empirical validation. In doing so, the paper shifts the discussion away from what students should not do with AI and towards the capabilities that higher education ought to cultivate deliberately.

Keywords: human moat; generative artificial intelligence; higher education; cognitive offloading; human agency

INTRODUCTION

Tools such as ChatGPT can now draft an essay, summarise a set of readings, write working code or outline an argument within seconds. For higher education, this creates an uneasy tension between efficiency and the cognitive effort on which learning depends (Wang & Zhang, 2026). Students who hand their thinking over to AI may complete tasks more quickly, yet learn less in the process, a pattern that has been described variously as cognitive offloading, metacognitive laziness and deskilling (Y. Fan et al., 2025; Yang et al., 2026).

Research has responded quickly, although in a fragmented way. Separate bodies of work examine critical

thinking (C. Li et al., 2026; Rahman et al., 2025), metacognition (Dhurumraj & Labuschagne, 2026), learner agency (G. Wang et al., 2026), evaluative judgement (Bearman et al., 2024), epistemic vigilance (Zainuddin et al., 2026) and AI literacy (Yan et al., 2025). Each strand sheds light on part of the problem. What is still missing is an account that identifies, as a coherent whole, the human capabilities that universities need to protect, and that explains why their importance grows rather than diminishes as AI becomes more capable. The literature is also tilted towards harm. Of the 173 records analysed here, 139 (80%) describe erosion, dependence or offloading, whereas only 62 (36%) report an intervention, scaffold or design intended to protect human capability. The search terms, which included cognitive offloading and deskilling, account for part of this imbalance, but the relative scarcity of protection studies remains notable. Educators are consequently accumulating a long list of what AI might damage while receiving comparatively little guidance on what to defend.

In Malaysia the gap is wider still. Only one record in the corpus had a Malaysia-affiliated author (Zainuddin et al., 2026), and none of the abstracts addressed the Malaysian context. Policy, however, has moved ahead of research. The Ministry of Higher Education's 2024 guidelines on GenAI in teaching and learning warn that uncontrolled use may weaken students' writing, evaluation, problem-solving and critical-thinking skills (Kementerian Pendidikan Tinggi [KPT], 2024), and universities have since introduced their own procedures, including the expectation that students declare any AI use in their assignments (Bernama, 2026). These measures recognise the risk and regulate misuse, but they stop short of specifying which capabilities students should retain or how teaching might develop them. The present paper responds by proposing the concept of the Human Moat. In business strategy, a moat refers to a durable advantage that rivals find difficult to replicate; in the argument developed here, the rival is the machine. The paper contributes in four ways. It defines the Human Moat and grounds the concept in task-based theories of automation and in cognitive offloading theory. It derives five dimensions from recent research on GenAI in higher education and distinguishes the concept from adjacent constructs. It specifies the mechanism through which the moat is built or eroded and expresses that mechanism as seven testable propositions. Finally, it translates the framework into implications and a research agenda, using Malaysian higher education as an illustrative setting. To make the framework usable, it also proposes indicative indicators for each dimension, practical implementation guidance and a staged plan for empirical validation.

Approach And Framework Development

This is a conceptual paper that adopts a theory-synthesis approach, integrating existing constructs into a new framework rather than testing relationships empirically. To avoid relying on a selective reading of the field, the synthesis was anchored in a defined corpus of published work. In terms of Jaakkola's (2020) typology of conceptual articles, it is a theory synthesis: it integrates a fragmented body of research into a framework that offers a new, higher-level view of the phenomenon. The development procedure and the decision rules applied at each step are reported below so that the reasoning behind the framework can be examined and repeated. Records were retrieved from the Web of Science Core Collection and Scopus on 15 October 2026 using the search strings reported in Table 1. Both strings combined three concept blocks, covering generative AI, the human capabilities at stake and the higher-education setting, and no restrictions on language, document type or publication year were applied.

Table 1. Search strategy (search date: 15 October 2026).

Database	Field searched	Search string	Records
Web of Science Core Collection	Topic (TS: title, abstract, author keywords, Keywords Plus)	TS=(("generative AI" OR "generative artificial intelligence") AND ("human agency" OR "cognitive offloading" OR "deskilling" OR "human capabilities") AND ("higher education" OR universit*))	114
Scopus	Title, abstract, keywords	TITLE-ABS-KEY(("generative AI" OR "generative artificial intelligence") AND ("human agency" OR "cognitive offloading" OR "deskilling" OR "human capabilities") AND ("higher education" OR universit*))	164

The two exports were merged, and duplicates were removed first by DOI and then by normalised title, leaving 173 unique records published between 2024 and 2027. The five records dated 2027 had already appeared online but had been assigned to forthcoming journal issues or proceedings volumes. Journal articles formed the largest group, followed by reviews, conference papers and book chapters. Since the corpus was used to identify themes rather than to synthesise effects, no records were excluded.

The corpus informed the framework in two ways. Titles, abstracts and keywords were first screened for terms associated with human capabilities, among them critical thinking, judgement, epistemic vigilance, metacognition, agency, ownership, creativity and ethics, and themes that were both frequent and recognisably human capabilities were retained as candidate dimensions (Table 2). In a second step, 146 passages describing what learners should retain, preserve or refrain from delegating were extracted and read closely in order to define each dimension and the mechanism connecting them.

Both steps were carried out with a script so that the counts can be reproduced. For each record, the title, abstract, author keywords and index keywords (Keywords Plus in Web of Science) were combined and converted to lower case. A record was counted under a theme if it contained at least one term from that theme's keyword set (Appendix A), with truncated stems such as creativ* capturing related word forms. Themes were not mutually exclusive, so a record could be counted under several of them, and ambiguous records were not reclassified by hand. The resulting figures should therefore be read as indicators of attention rather than as precise prevalence estimates. The same procedure, using separate keyword sets, was applied to classify records as describing erosion or protection. For the passage extraction, each abstract was split into sentences, and sentences matching a set of retention-related expressions (Appendix A) were extracted and then read in full. The underlying dataset, scripts and screening outputs have been retained and are available for verification.

This screening was descriptive in nature. It shaped the framework without constituting a test of it, and keyword frequencies indicate where scholarly attention is concentrated rather than what is effective. A full bibliometric and topic-modelling analysis of the corpus will be reported separately.

Framework development procedure. The framework was developed in six steps.

Step 1. Problem framing. The starting point was the gap identified in Section 1: research on GenAI in higher education is fragmented and oriented towards erosion, and it does not specify which human capabilities ought to be protected.

Step 2. Corpus construction. Records were retrieved from Web of Science and Scopus and de-duplicated, giving 173 unique records (Table 1).

Step 3. Theme screening. Titles, abstracts and keywords were screened for capability-related terms to identify candidate themes (Table 2; Appendix A).

Step 4. Dimension selection. As described above, themes that were both frequent and recognisably human capabilities were retained. Stated explicitly, the decisions in Table 2 reflect four criteria: (a) *salience*, with every retained theme appearing in at least 17% of records; (b) *capability rather than tool skill*, meaning it describes something learners do with their own minds rather than a skill for operating AI; (c) *delegability*, meaning GenAI can plausibly perform it on the learner's behalf, so that it is at risk of erosion; and (d) *theoretical warrant*, meaning its retention can be explained by task-based automation theory or cognitive offloading theory (Section 3.4). Themes meeting all four criteria were kept, conceptually overlapping themes were merged, and themes failing criterion (b) were reclassified as enabling conditions.

Step 5. Dimension definition. The 146 retention passages were read in full to define each dimension, identify its moat-eroding signal and select anchor sources (Table 4).

Step 6. Mechanism and propositions. The relationships described in the passages were combined with the

two theories to specify the enabling condition, the pedagogical switch, the two pathways and the seven propositions (Sections 4.3 to 4.5).

Table 2. Human-capability themes in the corpus (title, abstract and keywords). Records can match more than one theme.

Theme	Records	Share of 173	Decision
Ethics, integrity, responsibility	99	57%	Kept, merged with agency
Agency, ownership, autonomy	86	50%	Kept
AI literacy, prompting	76	44%	Treated as enabler
Critical thinking, higher-order thinking	74	43%	Kept, merged with evaluative judgement
Creativity, originality	60	35%	Kept
Metacognition, self-regulated learning	56	32%	Kept
Epistemic vigilance, verification, trust	38	22%	Kept
Evaluative judgement	29	17%	Merged with critical thinking

Note. Decisions apply the criteria in Step 4. All retained themes appeared in at least 17% of records. Ethics was merged with agency and evaluative judgement with critical thinking because each pair overlapped conceptually. AI literacy met the salience criterion but failed criterion (b), since it concerns the use of AI rather than a capability retained from it, and it was therefore treated as an enabling condition.

CONCEPTUAL BACKGROUND

Cognitive offloading: neither good nor bad by itself

Cognitive offloading refers to the use of an external tool to reduce mental effort. GenAI extends what can be offloaded beyond memory to reasoning, structuring and even judgement, and the available evidence suggests that the consequences cut both ways. Strategic offloading can release cognitive resources for higher-order reflection (Wang & Zhang, 2026), and the cognitive relief it provides has been linked to more positive attitudes and deeper engagement (W. Fan et al., 2026). Habitual offloading, on the other hand, is associated with weaker critical thinking, effort avoidance and what Hamad et al. (2026) term cognitive atrophy, as well as with a decline in learning accountability as students transfer epistemic trust to AI (Y. Liu et al., 2026).

Recent reviews arrive at a similar position. G. Wang et al. (2026) distinguish agency-supportive from agency-eroding patterns of use, and Suazo Galdames et al. (2026) conclude that although AI can support reflection, uncritical reliance tends to coincide with automation bias and superficial learning. Yang et al. (2026), for their part, show that students follow one of three pathways, namely deskilling, reskilling or upskilling. The pertinent question is therefore less whether students offload than what they must continue to do for themselves.

Existing Constructs and Their Limits

Several established constructs already address aspects of this question. Evaluative judgement, understood as the capability to judge the quality of one's own and others' work, is described by Bearman et al. (2024) as a uniquely human capability in the age of GenAI. Human or learner agency, which concerns students' control over their own learning, is the capability theme that appears most often in the corpus and is frequently framed in terms of human oversight (Bozkurt, 2025; Bozkurt et al., 2024). AI literacy encompasses the

knowledge and skills needed to use AI well, including critical evaluation and human agency (Tremblay et al., 2025; Yan et al., 2025). The construct closest to the present proposal is cognitive AI safety, for which Zainuddin et al. (2026) identify three factors: epistemic vigilance, cognitive agency and failure resilience.

These constructs share two limitations. Most are framed around risk or skill acquisition, so they explain what can go wrong or what needs to be learned while saying little about why certain human capabilities retain lasting value. They are also studied largely in isolation, with the result that educators receive a series of partial answers instead of a coherent target.

Why a "moat"

In strategic management, an economic moat denotes a durable advantage that shields a firm from competition. Applied to learning, the metaphor carries three claims that existing constructs do not make together. The first is that value is relative: a capability matters more when AI cannot easily reproduce it, so the human capabilities that remain are expected to gain in value as AI improves at routine work. The second is that the moat requires maintenance because, like its physical counterpart, it silts up when neglected, and capabilities that go unused gradually weaken. The third is that the moat is a shared asset, belonging to the learner yet built through the curriculum, pedagogy and assessment that institutions provide. Seen in this way, the Human Moat draws existing constructs together under a single value-based question, namely which human capabilities higher education should deliberately defend.

Theoretical Foundations

Two established theories help to explain why such a moat exists and how it changes over time.

The first is the task-based theory of automation. Autor et al. (2003) showed that computers substitute for routine tasks that follow explicit rules while complementing non-routine analytical and interpersonal tasks, so that the value of non-routine capabilities rises as routine work is automated. GenAI moves this boundary, since it can now perform some tasks previously regarded as non-routine, including drafting, summarising and generating ideas. The Human Moat comprises the capabilities that remain on the complementary side of the shifting boundary: judging quality, verifying claims, regulating one's own thinking, taking responsibility and producing original ideas. This theory supports the first claim of the metaphor, that the value of the moat is relative and increases with AI capability.

The second is cognitive offloading theory. Risko and Gilbert (2016) define cognitive offloading as the use of physical action or external tools to reduce the cognitive demands of a task. Offloading frequently improves immediate performance, but decisions to offload depend on metacognitive judgements about one's own ability and the effort involved, and such judgements are often poorly calibrated. The theory therefore supports the second claim, that the moat must be maintained, because capabilities that learners offload by default are no longer practised. It also explains why metacognitive self-regulation warrants treatment as a dimension in its own right, given that it governs every other decision about what to offload.

In combination, the two theories supply the mechanism behind the Human Moat. Task-based theory indicates what is worth retaining, while cognitive offloading theory accounts for how it is retained or lost.

How the Human Moat differs from adjacent constructs

Table 3. The Human Moat compared with adjacent constructs.

Construct	Core question	Main focus	What the Human Moat adds
Evaluative judgement (Bearman et al., 2024)	Can the learner judge the quality of work?	One capability	Places judgement within a wider set and explains why its value rises
Human or learner	Who controls the	Control and autonomy	Adds accountability and

agency (G. Wang et al., 2026; Bozkurt, 2025)	learning?		names the capabilities through which agency is exercised
AI literacy (Tremblay et al., 2025; Yan et al., 2025)	Can the learner use AI well?	Skills for using AI	Treats AI literacy as an enabler, not the goal; the goal is what the learner retains
Cognitive AI safety (Zainuddin et al., 2026)	Is the learner protected from AI-related cognitive risk?	Risk prevention	Shifts from avoiding harm to building value; adds creativity, accountability and a pedagogical switch
21st-century skills / 4Cs (Trilling & Fadel, 2009)	What skills do learners need for a changing world?	General competences, including collaboration and communication	Selects only the capabilities GenAI can take over and explains why their value rises
OECD Learning Compass 2030 (OECD, 2019)	Which competences help learners shape their future?	Student agency and transformative competences	Specifies the AI-related threat to agency and the pedagogical switch that protects it
UNESCO AI competency framework for students (UNESCO, 2024)	What should students know and be able to do about AI?	Human-centred mindset, AI ethics, AI techniques and system design	Focuses on what students keep as their own rather than on competence with AI
Human Moat	**Which human capabilities should higher education deliberately defend, and why?**	**A protected set of five capabilities**	**Integrates the above under one value-based rationale, with a mechanism and testable propositions**

Of the constructs compared in Table 3, cognitive AI safety is the closest. Its three factors (epistemic vigilance, cognitive agency and failure resilience) overlap with the second and fourth dimensions of the Human Moat, yet the two frameworks are oriented differently. Cognitive AI safety asks how learners can avoid harm, whereas the Human Moat asks what learners should retain because its value grows as AI improves. The two are best regarded as complementary, and the safety factors could serve as indicators within a future measure of the moat.

Table 3 also includes broader human-skills frameworks, which differ from the Human Moat in their selection logic. Frameworks such as the 4Cs of 21st-century learning (Trilling & Fadel, 2009) and the OECD Learning Compass 2030 (OECD, 2019) describe competences that are desirable in general, including collaboration, communication and social and emotional skills, without asking which of them a particular technology places at risk. The UNESCO AI competency framework for students (UNESCO, 2024) is AI-specific and includes a human-centred mindset, but, like AI literacy, it is organised around understanding, using and designing AI. The Human Moat applies a narrower test: a capability is included only if GenAI can plausibly perform it for the learner and if its value rises as GenAI improves. In practice, this means the framework can sit alongside the graduate attributes that universities already use, helping them decide which of those attributes need deliberate protection in AI-integrated courses.

The Human Moat Framework

Definition

The Human Moat is the set of distinctively human capabilities that learners retain as their own rather than delegate to generative AI, and whose value increases as AI absorbs routine cognitive work. It is built when AI augments human thinking and eroded when AI substitutes for it.

Five dimensions

Table 4. The five dimensions of the Human Moat.

Dimension	What the learner keeps	Moat-eroding signal	Anchor sources
1. Critical and evaluative judgement	Reasoning about AI output and judging its quality against standards	Accepting fluent output without analysis	Bearman et al. (2024); Vendrell & Johnston (2026); M. Fan & Chang (2026)
2. Epistemic vigilance	Verifying claims, detecting errors and calibrating trust in AI	Transferring epistemic authority to AI	Zainuddin et al. (2026); Ming & Li (2026); Y. Liu et al. (2026)
3. Metacognitive self-regulation	Planning, monitoring and deciding when to offload; keeping productive struggle	"Metacognitive laziness" and offloading by default	Y. Fan et al. (2025); Morales (2026); Almudara et al. (2026)
4. Agency and ethical accountability	Ownership of the work and responsibility for its consequences	Outsourced responsibility and weak psychological ownership	X. Li et al. (2026); Du & Yuan (2026); Wu & Lu (2026)
5. Creative generation	Producing original ideas, not only choosing among AI suggestions	Dependence that narrows originality	W. Li et al. (2026); Sobrinho (2026); Peters & Angelov (2025)

Critical and evaluative judgement forms the core of the moat. Critical thinking is the outcome most frequently studied in the corpus, and evaluative judgement gives it a concrete reference point in the learner's sense of what good work looks like. A student who cannot judge quality is poorly placed to use AI productively. Moreover, M. Fan and Chang (2026) found that simply instructing students to evaluate AI output did not lead them to do so, which suggests that this capability has to be designed into tasks rather than assumed.

Epistemic vigilance is the dimension most specific to AI. Because GenAI produces fluent and confident text that may nonetheless be wrong or fabricated, learners need to verify claims and calibrate their trust accordingly. Ming and Li (2026) observed that encountering such errors can prompt what they call algorithmic discernment, turning passive trust into active checking. Where vigilance is absent, students tend to shift epistemic trust to AI and to feel less accountable for their own learning (Y. Liu et al., 2026).

Metacognitive self-regulation is likely to shape whether offloading helps or harms. Morales (2026) distinguishes assistance that follows a learner's own attempt at recall, which preserves retrieval, from assistance that replaces that attempt and thereby displaces it. The capability at stake lies in judging when to persist with a difficult task and when delegation is appropriate.

Agency and ethical accountability provide perhaps the strongest argument for the moat, since AI can produce work but cannot be held accountable for it. Students who plan before turning to AI, evaluate its output and revise in their own voice retain psychological ownership of their work (X. Li et al., 2026), a stance that Du and Yuan (2026) describe as retained epistemic responsibility. In the corpus, agency and ethics were also the most closely linked pair of themes.

Creative generation safeguards originality. AI can broaden the range of ideas available to a learner, yet dependence on it has been associated with lower perceived creativity (W. Li et al., 2026). Sobrinho (2026) reports that human agency remains central at the beginning (preparation) and end (verification) of the creative process, even where AI shapes the intermediate stages.

Two capabilities that are prominent in human-skills frameworks, empathy and collaboration, are not treated as separate dimensions. Both matter for graduates, but they meet the selection criteria less clearly: relational capacities received too little support in the corpus to justify a separate dimension (Section 6.2), and they are

less directly open to delegation to GenAI than drafting, summarising or evaluating. Where collaboration was examined, the central concern was who retained agency in AI-supported group work (Wu & Lu, 2026), which the fourth dimension captures. For now, they are better seen as settings in which the moat capabilities are used, and they could become dimensions in their own right if future studies show students handing them over to AI.

Enabler and Switch

Two further elements complete the framework. AI literacy functions as an enabler rather than as a moat capability, because it concerns the effective use of AI rather than the retention of what AI cannot do; without it, however, learners struggle to tell augmentation from substitution (Yan et al., 2025; Zainuddin et al., 2026). Pedagogical design, in turn, is proposed to act as the switch. The same tool can build or erode the moat depending on how a task is structured. Scaffolded and structured use of AI has been found to support higher-order thinking (Hong et al., 2025; Vendrell & Johnston, 2026), whereas unstructured use appears to benefit mainly lower-order skills (Rahman et al., 2025).

Two Pathways

Accordingly, the framework distinguishes two pathways, illustrated in Figure 1. On the moat-building, or augmentative, pathway, the learner turns to AI only after making an independent attempt, then evaluates and verifies its output and takes ownership of the result. This route is associated with upskilling (Yang et al., 2026) and with perceiving AI as a complement to one's own thinking, which in turn predicts cognitive independence (Dominek et al., 2026). On the moat-eroding, or substitutive, pathway, AI replaces the learner's attempt and its output is accepted largely as given, a pattern that has been associated with dependence, deskilling and, in the longer term, cognitive atrophy (Hamad et al., 2026).

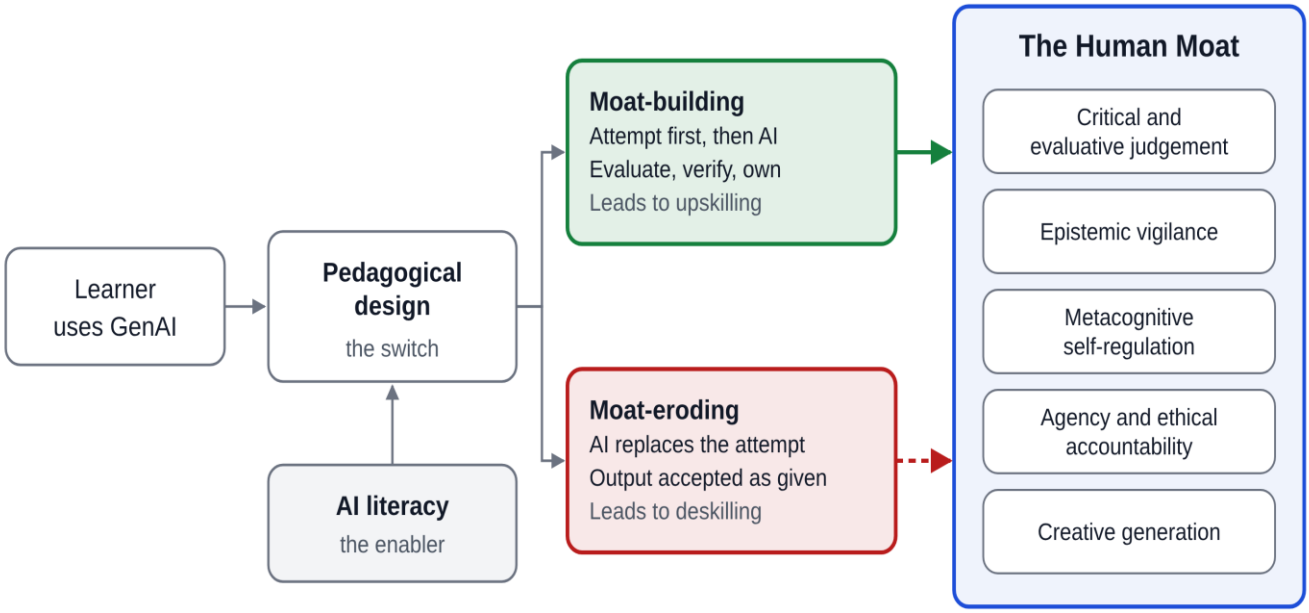

Figure 1. The Human Moat framework. Pedagogical design, supported by AI literacy, directs GenAI use towards either a moat-building or a moat-eroding pathway, which strengthens or weakens the five moat dimensions respectively.

Propositions

Seven propositions follow from the framework. The first four concern the two pathways and the factors that lead learners into one or the other, while the remaining three address the pedagogical switch, the mediating role of accountability and the longer-term value of the moat.

- **P1.** The effect of GenAI use on the moat depends on the mode of use: augmentative use is positively, and substitutive use negatively, associated with each moat dimension (Dominek et al., 2026; Yang et al., 2026).

- **P2.** Metacognitive self-regulation is positively associated with augmentative use and negatively associated with substitutive use (Y. Fan et al., 2025; Risko & Gilbert, 2016).

- **P3.** AI literacy moderates the relationship between the frequency of GenAI use and the moat dimensions, such that frequent use is more likely to strengthen the moat when AI literacy is high (Yan et al., 2025; Zainuddin et al., 2026).

- **P4.** Epistemic vigilance weakens the relationship between GenAI use and the transfer of epistemic trust to AI (Ming & Li, 2026; Y. Liu et al., 2026).

- **P5.** Tasks that require an independent attempt before AI is consulted (attempt-first designs) yield more augmentative use than tasks that permit AI from the outset (AI-first designs) (Morales, 2026; Vendrell & Johnston, 2026).

- **P6.** Agency and ethical accountability mediate the relationship between augmentative use and deep learning outcomes (X. Li et al., 2026; Y. Liu et al., 2026).

- **P7.** As GenAI capability increases, the value attached to moat capabilities in assessment and graduate employment is expected to increase (Autor et al., 2003).

Propositions 1 to 4 and 6 lend themselves to survey designs analysed through structural equation modelling. Proposition 5 calls for quasi-experimental work, and Proposition 7 requires longitudinal or labour-market evidence.

Indicative Indicators

For the framework to guide teaching and research, each dimension needs observable indicators. Table 5 proposes indicators that combine three kinds of evidence: self-report of dispositions and perceptions, behaviour (what learners actually do with AI output) and performance (the quality of assessed work). Behavioural and performance evidence are included for every dimension, since students tend to overestimate their own vigilance and independence when asked directly.

Table 5. Indicative indicators for the five dimensions of the Human Moat.

Dimension	Observable indicator	Instrument that could be adapted	Assessment evidence
1. Critical and evaluative judgement	Accuracy and justification when grading AI-generated work against a rubric; quality of reasons for accepting or rejecting AI suggestions	Critical-thinking disposition or skills measures; evaluative judgement tasks (Bearman et al., 2024)	Rubric-grading task with written justification
2. Epistemic vigilance	Proportion of AI claims checked against sources; number of seeded errors detected; calibration of trust in AI	Items developed from the epistemic vigilance factor of cognitive AI safety (Zainuddin et al., 2026)	Critique of AI output containing planted errors
3. Metacognitive self-regulation	Whether an independent attempt precedes AI use; accuracy of judgements about when to delegate	Metacognitive Awareness Inventory (Schraw & Dennison, 1994)	Process logs, planning notes and AI-use reflections

4. Agency and ethical accountability	Psychological ownership of the work; ability to explain and defend key decisions; transparent disclosure of AI use	Psychological ownership measures, as examined by X. Li et al. (2026)	Viva voce or oral defence; annotated AI-use declaration
5. Creative generation	Number and originality of ideas produced before AI use; extent to which AI suggestions are transformed rather than adopted	Kaufman Domains of Creativity Scale (Kaufman, 2012); expert rating of originality	Ideation before and after AI use; portfolio comparison

The indicators have not yet been validated. They are intended as a starting point for writing survey items and designing assessment tasks, and their validation is part of the research agenda in Section 6.3.

Implications For Malaysian Higher Education

Malaysian higher education offers an instructive case because national policy already acknowledges the risk. The Ministry of Higher Education's guidelines caution that uncontrolled GenAI use may weaken writing, evaluation, problem-solving and critical thinking, and they encourage students to practise fact-checking, decision-making and critical thinking. They also favour continuous assessment through presentations, oral questioning and viva voce, on the grounds that written work may no longer reflect a student's actual ability (KPT, 2024). Universities have developed their own procedures on this basis, and students are now expected to declare their use of AI (Bernama, 2026).

The Human Moat points to a further step, from regulating use to building capability. A declaration records that AI was used, but it reveals nothing about whether the student's judgement, vigilance or sense of ownership was strengthened or diminished in the process. The framework offers a way of giving the aims of the guidelines a structure that can be taught, assessed and measured.

Curriculum and Pedagogy

Three practices follow for curriculum and pedagogy. Tasks can be sequenced so that students draft, solve or recall before consulting AI, which is consistent with retrieval-preserving assistance (Morales, 2026). Courses can alternate AI-mediated and AI-free phases so that a degree of cognitive friction is retained (Vendrell & Johnston, 2026). Error detection can also be taught explicitly, for example by asking students to critique flawed AI output, given that instructions to evaluate are not sufficient on their own (M. Fan & Chang, 2026).

Because Malaysian programmes follow outcome-based education and must align with the learning outcome clusters of the Malaysian Qualifications Framework, second edition (2024), in force since 1 August 2026 (MQA, 2024), the dimensions can be integrated through existing course and programme learning outcomes rather than added as a separate course. Table 6 maps each dimension to the MQF clusters it most closely supports and suggests how a course learning outcome could be worded to make the GenAI context explicit.

Table 6. Mapping the Human Moat to MQF learning outcome clusters, with example course learning outcomes.

Dimension	MQF cluster(s)	Example course learning outcome
1. Critical and evaluative judgement	Cognitive skills	Evaluate the quality of AI-generated analysis against disciplinary standards and justify the evaluation.
2. Epistemic vigilance	Cognitive skills; digital skills	Verify claims in AI-generated content using authoritative sources and identify factual or reasoning errors.
3. Metacognitive self-regulation	Personal skills	Plan and monitor one's own learning, deciding appropriately when to use or withhold AI

		assistance.
4. Agency and ethical accountability	Leadership, autonomy and responsibility; ethics and professionalism	Take responsibility for AI-assisted work by disclosing AI use and defending key decisions orally.
5. Creative generation	Cognitive skills; entrepreneurial skills	Generate original solutions independently before using AI to extend or test them.

Assessment

Assessment can reinforce the same aims. Rather than judging only the final product, lecturers can ask students to show how they planned their work, which AI suggestions they accepted or rejected, and why. Evaluative judgement can be assessed directly by having students grade AI-generated work against a rubric and justify their grades (Bearman et al., 2024). Oral or reflective defence is also useful, since students who retain ownership are able to explain the decisions behind their work (X. Li et al., 2026), and this approach sits comfortably with the viva voce assessment that the national guidelines already favour.

Institutional Policy

At the institutional level, declarations of AI use could be extended into brief reflections in which students identify at least one judgement they made that AI did not. Staff development also deserves attention, because academics are themselves at risk of deskilling when they rely heavily on automated tools (Kostanek & Li, 2026); the moat applies to educators as much as to students. The framework also aligns with the national AI agenda. AI Malaysia Berhad was launched on 28 July 2026 as the permanent successor to the National AI Office, with a mandate that includes national AI capability development, and it now leads the National AI Action Plan 2026–2030 (AI Malaysia Berhad, n.d.). Graduate capabilities that complement AI, and not merely skills in operating it, could reasonably form part of that agenda.

Faculty Development

Whether the framework reaches the classroom depends largely on lecturers, since it is they who set the pedagogical switch. Design workshops would allow lecturers to take one of their existing assignments and rework it into an attempt-first task, adding an AI-use log and a short reflection. Grading calibration sessions, in which colleagues mark the same samples of AI-assisted work against a common rubric, would help lecturers sharpen the evaluative judgement they are asking students to develop. Over time, communities of practice within faculties could collect tasks that have worked well, which may also guard against the deskilling of educators themselves (Kostanek & Li, 2026). Most universities already have teaching and learning centres that could host such activities.

Phased Implementation

Institutions are likely to benefit from starting small. In a first semester, a few volunteer courses could adopt attempt-first tasks and at least one assessment aimed at a moat dimension, gathering baseline evidence with the indicators in Table 5. Designs that work could then be refined and shared over the following academic year, with course learning outcomes revised along the lines of Table 6. At programme level, the dimensions could eventually be mapped so that each is developed and assessed at least once in every year of study, with the evidence feeding into programme review and accreditation. The pace of this expansion is best set by what the earlier stages show.

Comparison With International Practice

Table 7 compares the framework with four widely cited international sources of guidance. There is considerable agreement on human oversight, on assessing the process of learning and on building staff capability. Where the sources differ from the present framework is in emphasis. Most international guidance concentrates on regulating use or redesigning assessment, while the Human Moat is concerned with the capabilities that such regulation and assessment are meant to protect, so the two can be used together. For

example, a task placed at a given level of the AI Assessment Scale (Perkins et al., 2024) could be paired with the Table 5 indicators that the task must still evidence.

Table 7. The Human Moat compared with international guidance on GenAI in higher education.

Source	Main emphasis	Convergence with the Human Moat	What the Human Moat adds
UNESCO Guidance for GenAI in education and research (UNESCO, 2023)	Human-centred, regulated adoption of GenAI	Human agency and oversight	Names the specific capabilities to protect and how pedagogy builds them
TEQSA, Assessment reform for the age of AI (Lodge et al., 2023)	Programme-wide assessment redesign; assessing the process of learning	Process-focused assessment and oral defence (Section 5.2)	Specifies which capabilities assessment should evidence
Russell Group principles on generative AI (Russell Group, 2023)	AI literacy for staff and students; adapting teaching, assessment and integrity policy	Staff development and assessment adaptation	Moves from AI literacy to retained human capability
AI Assessment Scale (Perkins et al., 2024)	Graduated levels of permitted AI use in assessment tasks	Explicit task design for AI use, similar to attempt-first designs	Links each level of permitted use to the capabilities the task must still demonstrate

Because the framework draws on general theories of automation and offloading, and is consistent with guidance issued in several countries, there is little in it that is specific to Malaysia. The MQF mapping in Table 6 shows one way of adapting it to a national setting, and a similar exercise could be carried out with other qualifications frameworks.

Institutional and Cultural Challenges

Putting the framework into practice in Malaysia is unlikely to be straightforward. Where learning is examination-oriented and rewards final grades, students may be tempted to treat GenAI as a shortcut, so process-based assessment needs explicit weighting in course grades if students are to take it seriously. Malaysian society is characterised by high power distance (Hofstede et al., 2010), and students may be reluctant to question authoritative-sounding sources, including fluent AI output; lecturers may therefore need to model and explicitly reward the questioning that epistemic vigilance requires. Oral defence and process assessment are costly in large classes, which can be eased through sampling, short structured vivas and peer review. Readiness and access may vary between public and private institutions and between well-resourced and less well-resourced campuses, so task designs should not assume access to premium tools. Students who study in a second language may rely on GenAI for linguistic expression; course policies should distinguish language support, which may be legitimate, from the substitution of thinking.

DISCUSSION AND CONCLUSION

Contributions

The paper makes three theoretical contributions. It integrates five constructs that are usually examined separately into a single framework, with a shared rationale for their importance. It connects research on GenAI in education with task-based theories of automation, offering an explanation of why the value of certain human capabilities rises rather than falls as AI advances. It also specifies a mechanism, in the form

of a pedagogical switch between augmentative and substitutive use, and states that mechanism as testable propositions.

For practitioners, the framework provides a positive target. Instead of asking only what students must avoid doing with AI, institutions can ask which capabilities each course develops and how assessment demonstrates that development. The indicators, curriculum mapping, faculty development activities and phased implementation outlined in Sections 4.6 and 5 make that target operational.

LIMITATIONS

Several limitations should be acknowledged. The framework is conceptual and has yet to be tested empirically. The corpus spans a short and rapidly changing period (2024–2027), and keyword screening indicates where research attention is concentrated rather than what is effective. Because the search terms favoured erosion-related concepts such as cognitive offloading and deskilling, studies of harm may be over-represented. The propositions rest largely on cross-sectional and qualitative evidence, almost all of it from outside Southeast Asia, so their applicability to Malaysian settings remains to be established. Relational capacities such as empathy and mentorship were raised in some studies (Bozkurt et al., 2024; Saravanan et al., 2026) but received too little support in the corpus to justify a separate dimension, although they may warrant one as evidence accumulates. In addition, the indicators in Table 5 and the mapping in Table 6 are proposals derived from the literature and have not yet been validated, and the comparison with international guidance in Table 7 is illustrative rather than systematic.

Research agenda

Several directions for further research follow. A full bibliometric and topic-modelling analysis could establish whether the five dimensions emerge as distinct clusters and could quantify the gap between studies of erosion and studies of protection. A Human Moat Index could be developed and validated among Malaysian university students, allowing Propositions 1 to 4 and 6 to be tested through structural equation modelling. Quasi-experimental comparisons of attempt-first and AI-first task designs in Malaysian classrooms would address Proposition 5, while longitudinal work could examine whether assessment systems and employers come to place greater weight on moat capabilities, as Proposition 7 predicts. The framework could also be extended to educators by examining how lecturers' own moat capabilities change as they adopt GenAI.

Empirical validation of the framework is proposed in four stages. The first would be expert content validation, in which a panel of academics, educational developers and quality assurance officers from Malaysian public and private universities would rate the relevance and clarity of the dimensions and indicators, using the content validity index (Lynn, 1986) or a Delphi procedure. Focus groups and interviews with students and lecturers would examine whether the dimensions are recognisable in practice and whether empathy or collaboration should be added. Items for a Human Moat Index would then be generated from Table 5, piloted and examined through exploratory and confirmatory factor analysis. Finally, the validated index would be used to test Propositions 1 to 4 and 6 through structural equation modelling, while pilot studies of attempt-first designs in Malaysian classrooms would test Proposition 5.

CONCLUSION

Generative AI will continue to take on routine cognitive work, and higher education therefore needs to be clear about what it is defending as well as what it is restricting. The Human Moat identifies five capabilities, namely critical and evaluative judgement, epistemic vigilance, metacognitive self-regulation, agency and ethical accountability, and creative generation, and it proposes that pedagogy largely shapes whether they are strengthened or eroded. For Malaysian universities, which already have integrity rules in place, the framework suggests a practical next step: designing teaching and assessment so that graduates retain the capabilities that make them valuable alongside AI.

Ethical Approval

Not applicable. This conceptual study analysed published bibliographic records and did not involve human participants or animals.

Conflict of Interest

The authors declare that they have no conflicts of interest regarding the publication of this study.

Data Availability Statement

The bibliographic records analysed in this study were exported from Web of Science and Scopus, which are subscription databases. The search strings in Table 1 allow the corpus to be reproduced. The merged record list is available from the corresponding author on reasonable request.

REFERENCES

1. AI Malaysia Berhad. (n.d.). AI Malaysia Berhad (National AI Office). Retrieved October 19, 2026, from https://ai.gov.my/faq/ai-malaysia-berhad/
2. Almudara, S. B., Alshehri, O. A., Alzahrani, S. Y., Abdellatif, M. S., Idris, A. I., & Ibrahim, A. R. (2026). Working memory and generative AI tools in higher education: A systematic review. Research Journal in Advanced Humanities, 7(2). https://doi.org/10.58256/ca66cf85
3. Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly Journal of Economics, 118(4), 1279–1333. https://doi.org/10.1162/003355303322552801
4. Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. Assessment & Evaluation in Higher Education, 49(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321
5. Bernama. (2026, July 22). Generative AI guidelines help universities manage technology use. The Sun. https://thesun.my/news/malaysia-news/people-issues/generative-ai-guidelines-help-universities-manage-technology-use/
6. Bozkurt, A. (2025). The three laws of artificial intelligence: Re-evaluating human-AI agency and interaction in a time of the generative and agentic AI Ren[ai]ssance. Open Praxis, 17(3), 421–428. https://doi.org/10.55982/openpraxis.17.3.794
7. Bozkurt, A., Xiao, J., Farrow, R., Bai, J. Y. H., Nerantzi, C., Moore, S., Dron, J., Stracke, C. M., Singh, L., Crompton, H., Koutropoulos, A., Terentev, E., Pazurek, A., Nichols, M., Sidorkin, A. M., Costello, E., Watson, S., Mulligan, D., Honeychurch, S., … Asino, T. I. (2024). The manifesto for teaching and learning in a time of generative AI: A critical collective stance to better navigate the future. Open Praxis, 16(4), 487–513. https://doi.org/10.55982/openpraxis.16.4.777
8. Dhurumraj, T., & Labuschagne, C. (2026). Artificial intelligence in higher education: Reconfiguring metacognition in unequal contexts. Frontiers in Education, 11, Article 1918374. https://doi.org/10.3389/feduc.2026.1918374
9. Dominek, D. L., Kapusi, V., Ceglédi, S., & Szűts, Z. (2026). AI as cognitive complement or replacement? Perceived AI role and cognitive independence among university students. Computers, 15(8), Article 507. https://doi.org/10.3390/computers15080507
10. Du, Y., & Yuan, Y. (2026). Epistemic dependence in AI-mediated learning within higher education: A framework for student judgement and responsibility. Assessment & Evaluation in Higher Education. Advance online publication. https://doi.org/10.1080/02602938.2026.2732250
11. Fan, M., & Chang, P. (2026). Telling students to evaluate does not make it happen: Task stage, offloading tendency, and error detection in AI-assisted student writing. Behavioral Sciences, 16(9), Article 1671. https://doi.org/10.3390/bs16091671
12. Fan, W., Cheng, L., Wang, Y., Zhao, Q., & Li, Y. (2026). In-class AI use and attitudes among university students: The different mediating roles of cognitive relief and cognitive offloading. Behavioral Sciences, 16(6), Article 1014. https://doi.org/10.3390/bs16061014

13. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544
14. Hamad, W., Deng, H., & Hossain, M. (2026). The dark side of generative artificial intelligence in personal learning: A cognitive perspective. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2026).
15. Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations: Software of the mind (3rd ed.). McGraw-Hill.
16. Hong, H., Vate-U-Lan, P., & Viriyavejakul, C. (2025). Cognitive offload instruction with generative AI: A quasi-experimental study on critical thinking gains in English writing. Forum for Linguistic Studies, 7(7), 325–334. https://doi.org/10.30564/fls.v7i7.10072
17. Jaakkola, E. (2020). Designing conceptual articles: Four approaches. AMS Review, 10(1), 18–26. https://doi.org/10.1007/s13162-020-00161-0
18. Kaufman, J. C. (2012). Counting the muses: Development of the Kaufman Domains of Creativity Scale (K-DOCS). Psychology of Aesthetics, Creativity, and the Arts, 6(4), 298–308. https://doi.org/10.1037/a0029751
19. Kementerian Pendidikan Tinggi. (2024). Garis panduan penggunaan teknologi kecerdasan buatan generatif (KBG) dalam pengajaran dan pembelajaran (PdP) pendidikan tinggi [Guidelines on the use of generative artificial intelligence technology in teaching and learning in higher education]. Jabatan Pendidikan Tinggi. https://repositori.mohe.gov.my/109/
20. Kostanek, E., & Li, F. (2026). The false sense of teaching: Educator deskilling in the age of AI. In Overdependence on AI in academia (pp. 289–316). IGI Global. https://doi.org/10.4018/979-8-3373-7097-2.ch010
21. Li, C., Cui, H., & Hagedorn, L. S. (2026). The cognitive impact of ChatGPT in higher education: A systematic review of critical and creative thinking outcomes. Computers and Education: Artificial Intelligence, 10, Article 100571. https://doi.org/10.1016/j.caeai.2026.100571
22. Li, W., Zhang, W., Li, M., & Zhang, Y. (2026). Generative AI as an external cognitive tool: Cognitive dependence, AI literacy, and perceived academic creativity. Journal of Intelligence, 14(8), Article 185. https://doi.org/10.3390/jintelligence14080185
23. Li, X., Xian, X., & Du, P. (2026). From AI dependence to reflective collaboration: Psychological ownership, competency anxiety, and perceived support in AI-assisted learning. Frontiers in Psychology, 17, Article 1903784. https://doi.org/10.3389/fpsyg.2026.1903784
24. Liu, Y., Jou, M., & Hao, Y. (2026). From cognitive offloading to learning accountability erosion: How generative AI reshapes students' epistemic trust and educational engagement. Interactive Learning Environments. Advance online publication. https://doi.org/10.1080/10494820.2026.2680276
25. Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). Assessment reform for the age of artificial intelligence. Tertiary Education Quality and Standards Agency.
26. Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research, 35(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017
27. Malaysian Qualifications Agency. (2024). Malaysian Qualifications Framework (MQF) (2nd ed., 2024). Malaysian Qualifications Agency.
28. Ming, W., & Li, Q. (2026). From cognitive offloading to algorithmic discernment: A grounded theory inquiry into generative AI's shifting impact on university students' critical thinking. Interactive Technology and Smart Education. Advance online publication. https://doi.org/10.1108/ITSE-05-2026-0199
29. Morales, C. (2026). Retrieval interruption framework: AI-assisted cognition and retrieval-dependent learning in higher education. Education Sciences, 16(8), Article 1179. https://doi.org/10.3390/educsci16081179
30. OECD. (2019). OECD future of education and skills 2030: OECD learning compass 2030. OECD.
31. Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. Journal of University Teaching and Learning Practice, 21(6). https://doi.org/10.53761/q3azde36

32. Peters, M., & Angelov, D. (2025). Redefining assessment tasks to promote students' creativity and integrity in the age of generative artificial intelligence. International Journal for Educational Integrity, 21(1), Article 25. https://doi.org/10.1007/s40979-025-00201-x
33. Rahman, G., Almutairi, E. A. A., Almutairi, M., & Mudhsh, B. A. (2025). The impact of AI-powered text generation on undergraduate students' critical thinking: A study on ChatGPT usage. Journal of University Teaching and Learning Practice, 22(7). https://doi.org/10.53761/srkt0510
34. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002
35. Russell Group. (2023). Russell Group principles on the use of generative AI tools in education. Russell Group.
36. Saravanan, P., Meissner, C., & Glenn, J. (2026). 'It's here and we can't fight it': Engineering faculty perceptions of generative AI, student use, and the future of engineering education. Frontiers in Education, 11, Article 1880863. https://doi.org/10.3389/feduc.2026.1880863
37. Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19(4), 460–475. https://doi.org/10.1006/ceps.1994.1033
38. Sobrinho, A. B. F. (2026). Exploring how generative AI reshapes artistic creativity. Journal of Creative Behavior, 60(2), Article e70113. https://doi.org/10.1002/jocb.70113
39. Suazo Galdames, I. C., Santiesteban Velázquez, M., Saracostti, M., & Chaple Gil, A. M. (2026). Between cognitive offloading and critical autonomy: A systematic review of the epistemic implications of generative AI in higher education. Frontiers in Education, 11, Article 1878665. https://doi.org/10.3389/feduc.2026.1878665
40. Tremblay, C., Poellhuber, B., Sénécal, A.-M., & Roy, N. (2025). Design and validation of the psychometric properties of a questionnaire measuring AI literacy among university students: Preliminary results. In Artificial Intelligence in Education: AIED 2025 (Communications in Computer and Information Science, Vol. 2591, pp. 210–217). Springer. https://doi.org/10.1007/978-3-031-99264-3_26
41. Trilling, B., & Fadel, C. (2009). 21st century skills: Learning for life in our times. Jossey-Bass.
42. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO.
43. UNESCO. (2024). AI competency framework for students. UNESCO.
44. Vendrell, M., & Johnston, S.-K. (2026). Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education. Computers and Education: Artificial Intelligence, 10, Article 100572. https://doi.org/10.1016/j.caeai.2026.100572
45. Wang, G., Wang, W., Yang, D., & Ren, J. (2026). Generative AI, cognitive offloading, and learner agency in higher education: A scoping review. Behavioral Sciences, 16(7), Article 1150. https://doi.org/10.3390/bs16071150
46. Wang, S., & Zhang, H. (2026). Pedagogical partnerships with generative AI in higher education: How dual cognitive pathways paradoxically enable transformative learning. International Journal of Educational Technology in Higher Education, 23(1), Article 11. https://doi.org/10.1186/s41239-026-00585-x
47. Wu, Y., & Lu, X. (2026). A helping hand or a dominant partner? Individual perceptions of GenAI reliance and human agency in collaborative learning. British Journal of Educational Technology. Advance online publication. https://doi.org/10.1111/bjet.70090
48. Yan, W., Nakajima, T., & Sawada, R. (2025). Beyond tool use: Tracking the evolution of generative AI literacy among university students through a process-oriented investigation. Computers and Education: Artificial Intelligence, 9, Article 100465. https://doi.org/10.1016/j.caeai.2025.100465
49. Yang, B., Sun, Y., Zeng, Z., & Li, Q. (2026). Deskilling, reskilling, or upskilling? Unpacking the pathways of student adaptation to generative artificial intelligence. International Journal of Information Management, 87, Article 103002. https://doi.org/10.1016/j.ijinfomgt.2025.103002
50. Zainuddin, Z., Churiyah, M., Xiaoyu, W., & Halili, S. H. (2026). From AI literacy to cognitive AI safety: Rethinking safe AI use in education. Interactive Technology and Smart Education. Advance online publication. https://doi.org/10.1108/ITSE-12-2025-0396

APPENDIX A. SCREENING KEYWORD SETS

Table A1. Keyword sets used in the theme screening, the erosion and protection classification, and the passage extraction.

Theme or category	Keyword set (case-insensitive; * = truncated stem)
Ethics, integrity, responsibility	integrity; ethic*; responsib*
Agency, ownership, autonomy	agency; ownership; autonom*; authorship; authorial; independen*
AI literacy, prompting	AI literacy; prompt*; literac*
Critical thinking, higher-order thinking	critical thinking; higher-order; analytical; reasoning
Creativity, originality	creativ*; originality; innovat*
Metacognition, self-regulated learning	metacogni*; self-regulat*; self regulat*; monitor*
Epistemic vigilance, verification, trust	epistemic; vigilan*; verif*; trust calibrat*; automation bias; hallucinat*
Evaluative judgement	evaluative judg*; judgement; judgment; evaluat* (AI/GenAI) output; quality of work
Erosion	erod*; erosion; deskill*; atroph*; overrelian*; over-relian*; dependen*; offload*
Protection	interven*; scaffold*; design principle*; pedagogical design; experiment*
Retention passages	uniquely human; irreplaceab*; cannot be replaced, automated or delegated; human agency; human judg*; human oversight; human expertise; evaluative judg*; epistemic vigilance, agency or responsibility; preserv*; safeguard*; deskill*; erod*; ownership; distinctively human; higher-order; productive struggle; complement*