IIAR JOURNALS

Innovating Knowledge. Advancing Research.

IIAR Journal of Health, Medicine, and Biomedical Research

Vol 1, No 1, 2026

Medical Education in the Age of Generative AI: A Competency Framework for Clinical Reasoning, Verification and Accountability

IGWE Emmanuel Sunday[1], NWOGBOJI Emmanuel Usulor[2], BAMIGBOLA Paul Idowu[3], ATOLAGBE Umar Kawu[4], OGAH Sunday Ogbo[5]

[1,2,3,4,5] Veristzon Global Research Lab

ABSTRACT

Generative artificial intelligence (GenAI), the branch of artificial intelligence (AI) that creates new content, can answer many of the examination questions used to certify competence in medicine and nursing, yet unstructured access has not reliably made learners better clinicians. Existing competency frameworks address single professions, practising clinicians or faculty, and none treats clinical reasoning, verification and accountability as separately assessable competencies for both medical and nursing learners. The purpose of this theoretical paper is to develop such a framework, guided by three questions: which competencies GenAI makes more important for medical and nursing students; which of these may be practised with AI assistance and which must be certified unaided; and how such a framework can be implemented in African medical and nursing schools. The evidence base is an integrative review. The population was peer-reviewed journal literature on GenAI in medical, nursing and health professions education published from 2023 to 2026, and purposive, criterion-based sampling across three screening rounds in Europe PMC (MEDLINE), PubMed, ERIC, OpenAlex, Google Scholar and the Directory of Open Access Journals yielded a sample of 143 articles. Each article was verified against its Crossref publisher record and screened for retractions, corrections and expressions of concern. The sample included 15 randomised studies, 38 studies of nursing or nursing-inclusive populations and 21 studies from African settings or African-led teams. Findings were charted in a literature matrix, synthesised deductively by five domains and inductively by competency, and interpreted through cognitive offloading theory, automation bias, competency-based assessment and Rest's four-component model. The synthesis showed that unstructured GenAI use did not consistently improve reasoning; those trainees detected AI errors at near-chance levels, although brief instruction improved AI-assisted performance in small studies; and that dependency was associated with weaker metacognition and unverified acceptance of AI output. Building on the complementarity condition, that assistance augments performance only where the user can evaluate its output, the paper proposes the Reason–Verify–Account (RVA) Competency Framework. It comprises 12 competencies in three clusters: clinical reasoning that holds without the tool, verification of machine output, and accountability for decisions the learner did not generate. The framework applies a rule of assisted practice and unaided certification, sequences competencies across foundation, clinical and transition-to-practice phases, and sets out a low-cost implementation path for African schools. Schools should state for each assessment whether AI is permitted, teach verification early, require disclosure and human authorship, and train faculty first. Validation through expert consensus and longitudinal testing is recommended.

Keywords: medical education; nursing education; generative AI; clinical reasoning; competency framework

1. INTRODUCTION

Generative artificial intelligence (GenAI), the branch of artificial intelligence (AI) that creates new content, entered health professions education faster than guidance for its responsible use (Popa et al., 2026; Saloojee et al., 2026). Within months of its public release, ChatGPT performed at or near the passing threshold on all three steps of the United States Medical Licensing Examination without specialised training (Kung et al., 2023), and a later meta-analysis put the pooled accuracy of GPT-4 on licensing examinations at 81% (M. Liu et al., 2024). Much of what written examinations ask of medical and nursing students can now be answered by GenAI systems.

Student adoption of these tools has outpaced institutional responses. In Nigeria, 61.9% of medical, dental and allied health students had already used ChatGPT (Iliyasu et al., 2025), and a survey of 1,443 medical students in Sudan found that awareness and use varied with gender, institutional setting and infrastructure (Ahmed et al., 2025). The same students worried about dependency, inaccuracy and ethics (Iliyasu et al., 2025; Nantale et al., 2025). Institutions have moved more slowly. Most US osteopathic medical schools surveyed had no GenAI policy or training for students, faculty or administrators (Ichikawa et al., 2025), and in nursing, favourable individual attitudes sat alongside weak institutional readiness, few standardised curricula and limited faculty training (El-Banna et al., 2025).

Access alone, however, has not produced better clinicians. In a randomised trial, physicians given a large language model (LLM) reasoned no better diagnostically than colleagues using conventional resources, even though the LLM alone scored 16 percentage points higher than the conventional-resources group (Goh et al., 2024). The model therefore performed better when used alone than when used by the physicians to whom it was provided. A South African trial found a similar result at the bedside: final-year students who used GPT-4o during real-patient examinations scored no higher than controls, and 37% reported distraction (Saloojee et al., 2026). In trauma simulation, unstructured access did not improve decision-making and may have disrupted team reasoning (Gustafsson et al., 2026). Structured use has produced different results. Surgical residents who followed the ABCD framework ("articulate, brainstorm and benchmark, critique and customize, and decide and discuss") made better decisions than those who used the model without guidance (Lai et al., 2026), and nursing students trained to use ChatGPT within the nursing process showed gains in problem-solving and competency (Arkan et al., 2025). The manner of use, rather than access alone, appears to determine the outcome.

These findings raise a competency question that curricula have not yet answered. If a generative system can draft the clinical note, propose the diagnosis and summarise the guideline, what must a medical or nursing student still be able to do unaided, and how should schools establish that competence? A scoping review found no standard approach or consensus on AI competencies (Simoni et al., 2025), and evidence on measurable educational outcomes remains poor (Feigerlova et al., 2025). The frameworks published since 2024 make useful but partial contributions. These frameworks address clinicians in practice (Cao et al., 2026), postgraduate training (Kabbash et al., 2026) or faculty (Meyers & Miller, 2026), or are confined to nursing (Xie et al., 2026) or to medicine (Corral-Gudino et al., 2026). None of those reviewed treats clinical reasoning, verification and accountability as linked but separately assessable competencies across both professions, and none was designed for African schools, where formal AI education lags behind (Adefolalu, 2026).

This paper develops such a framework. The paper argues that GenAI makes three capacities more, not less, important for learners, and the inquiry is guided by three questions. (a) Which competencies does GenAI make more important for medical and nursing students? (b) Which of these may be practised and assessed with AI assistance, and which must be certified unaided? (c) How can the framework be implemented in African medical and nursing schools? Throughout, "learners" refers to medical and nursing students and to postgraduate trainees.

© IIARJournal. All rights reserved.

2. LITERATURE REVIEW

2.1 CONCEPTUAL REVIEW

The framework rests on five concepts, which are defined here before the theory and evidence are reviewed.

2.1.1 GenAI and large language models

GenAI refers to AI systems that create new content in response to prompts (Ichikawa et al., 2025); LLMs such as GPT-4 are the text-generating subset most used in education. LLM outputs are fluent and often correct, and that fluency is itself a source of risk, because an error presented in the register of expertise is more difficult to question than one expressed with uncertainty. Hallucination, the generation of confident-sounding but fabricated content (Jin et al., 2023), was among the concerns most often reported by learners in one systematic review (59%; Bodke & Jumle, 2026).

2.1.2 Clinical reasoning

Clinical reasoning, closely related to clinical decision-making, is the process by which a clinician gathers, evaluates and interprets diagnostic information (Brügge et al., 2024), generates and tests hypotheses, and reaches decisions under uncertainty. Its stages matter here because AI performance varies across these stages. Frontier LLMs were highly accurate on final diagnoses but performed poorly on differential diagnosis and on handling uncertainty (Rao et al., 2026). Learners must therefore retain ownership of the stages at which these models are weakest.

2.1.3 Verification

Verification is the act of checking machine output against evidence, the patient and the context before relying on it. Jin et al. (2023) proposed a "retrieve, summarise and verify" paradigm in which the user retains full responsibility for checking LLM output. In this paper, verification covers accuracy, consistency, fit to the patient and local context, and bias; bias is included because GPT-4 has been shown to produce vignettes and differential diagnoses that stereotype race, ethnicity and gender (Zack et al., 2024).

2.1.4 Accountability

Accountability is the obligation to answer for a decision, including one the learner did not generate but will sign. A scoping review identified "accountability attribution" as a distinct ethical dimension of AI in medical education (Li et al., 2025). In nursing, professionalism with AI has been framed as a process in which learners "disclose, verify, and revise under human authorship" (Ali & Shaban, 2026). Accountability therefore has an individual dimension (ownership, disclosure and integrity) and an institutional dimension (policy and governance).

2.1.5 Cognitive offloading, deskilling and never-skilling

Cognitive offloading is the use of physical action, including external tools, to alter the information-processing requirements of a task so as to reduce cognitive demand (Risko & Gilbert, 2016). Three failure modes have been described when offloading becomes habitual in clinical education (Corral-Gudino et al., 2026; Y. Liu et al., 2026; Turney et al., 2026). *Deskilling* is the loss of skills already held; *never-skilling*, also termed upskilling inhibition, is the failure to acquire such skills; and *mis-skilling* is the acquisition of incorrect habits. Xu et al. (2026) describe the specific risk of "narrative cognitive substitution", in which pre-organised AI narratives shift clinicians from active hypothesis generation to passive validation.

2.2 THEORETICAL REVIEW

Four theoretical lenses inform the framework. Each accounts for a different part of the risk that GenAI poses to learners, and together the four lenses explain why access alone has yielded limited benefit.

Cognitive offloading theory. Risko and Gilbert (2016) place offloading within a metacognitive framework that explains what triggers it and examines its cognitive consequences. Extending that view, this paper argues that offloading is efficient when the aid is reliable and the user can recognise when it is not, but harmful when it displaces the processing through which skill is built. For learners, the harmful case is the norm, because the skill being offloaded is the very skill being acquired. This asymmetry between expert and novice is what the

© IIARJournal. All rights reserved.

reasoning competencies of the framework are designed to address.

Automation bias. Parasuraman and Manzey (2010) explain complacency and automation bias as attentional effects: users give automated advice too much weight and monitor it too little, especially under load. Clinical experiments confirm the effect. Radiologists at every experience level were swayed by incorrect AI suggestions in mammography, and inexperienced readers most of all (Dratsch et al., 2023). Learners, as the least experienced users, are likely to be among the most exposed. Parasuraman and Manzey's review also found that training or instructions alone do not prevent automation bias. If the same holds for learners, verification cannot be treated as a single instructional event; it must be built into task design and assessment, which is the premise of the verification cluster.

Competency-based assessment. Miller's (1990) pyramid distinguishes four levels of clinical competence (knows, knows how, shows how and does) and holds that assessment must reach the level at which performance matters. GenAI weakens assessment at the lower levels, since a model can supply much of the "knows" and "knows how" on demand. Recent work extends this logic to entrustment, classifying GenAI uses by educational stakes and AI autonomy (Patil et al., 2026). The pyramid supplies the logic of the assessment rule in Section 4.1.3: the higher the stakes, the more the evidence of competence must come from the learner alone.

Rest's four-component model. In Rest's model, moral action requires moral sensitivity, judgement, motivation and character, and nursing scholars have used it to trace how AI affects each component, including motivation through over-reliance (Sengul et al., 2025). Its value for this paper is that it treats accountability as a set of capacities that can be developed, rather than as a rule to be obeyed.

From these lenses the paper derives one integrating proposition, the *complementarity condition*: GenAI assistance augments human performance only where the user can independently evaluate its output. Where the user cannot, assistance substitutes for competence rather than extending it, and machine errors pass through unchecked. The condition carries a significant implication: the learners who most need assistance are the least able to judge its soundness. The framework is built to secure this condition.

2.3 EMPIRICAL REVIEW

The empirical evidence is organised by the five domains used in the search (Section 3.4), followed by the evidence from African settings.

2.3.1 GenAI and clinical reasoning

The trial evidence falls into two broadly consistent groups. Where access was unstructured, results were null or negative: no benefit was observed among practising physicians (Goh et al., 2024) or among final-year students examining patients at the bedside in South Africa (Saloojee et al., 2026), and team reasoning in trauma simulation may have deteriorated (Gustafsson et al., 2026). Structured use, by contrast, was associated with benefit. LLM-simulated history taking with AI-generated structured feedback improved medical students' clinical decision-making (Brügge et al., 2024), and GPT-simulated patients produced higher history-taking scores than role-play (Z. Wang et al., 2025). One trial, however, departed from this pattern: medical students given ChatGPT-4o for rheumatology vignettes identified the correct diagnosis more often than controls (77.5% vs. 32.4%) and outperformed the model alone, although with longer completion times (Roemer et al., 2026). This exception is informative. Naming a single correct diagnosis is the step at which LLMs are strongest (Rao et al., 2026), whereas the null trials assessed broader diagnostic reasoning or performance with real patients and teams. These trials are also small; five enrolled between 21 and 73 participants (Brügge et al., 2024; Goh et al., 2024; Gustafsson et al., 2026; Roemer et al., 2026; Saloojee et al., 2026). A meta-analysis of 20 randomised trials found short-term gains in knowledge and clinical skills with GenAI-assisted teaching but smaller gains at follow-up (Table 1), and its authors concluded that GenAI may supplement, but should not replace, conventional instruction (C. Wang et al., 2026).

2.3.2 Verification of AI output

The evidence on verification is less reassuring. General practice trainees identified ChatGPT-4o hallucinations with only 55.0% accuracy, close to chance (Zhou et al., 2025). Model accuracy itself fell

© IIARJournal. All rights reserved.

where clinical reasoning was hardest: GPT-4 averaged 86% on licensing-style items but 61–71% on multi-step vignettes and 48–63% on image items (Bodke & Jumle, 2026). Errors are therefore most likely in the tasks that learners find hardest to check. Context and language add to the problem. Without context, ChatGPT failed a French-language pulmonology examination, and it struggled with synthesis, clinical judgement and non-English items (Cherif et al., 2024). Some verification behaviours do respond to brief instruction. A 10-minute tutorial raised radiology residents' LLM-assisted diagnostic accuracy above both untutored LLM use and internet search (Kim et al., 2025), and exposure to the AI's flaws increased first-year students' critical stance toward AI content (Çiçek et al., 2025). The tutorial study involved only nine residents, however, so it is better read as a reason to test brief instruction than as evidence that it works.

2.3.3 Accountability and academic integrity

The evidence in this domain is dominated by policy gaps and learner uncertainty rather than by outcome studies. Nursing students distinguished ethical from unethical uses with 93% accuracy but were unsure where AI content becomes dishonesty (Dodson et al., 2025); knowledge of the rules therefore did not extend to the location of the boundary. In Sudan, students' ethical awareness was strongest for privacy and accountability but had been acquired mainly through informal learning (Muhalil et al., 2026). Two recent studies suggest that the gap can be closed. A student-partnered Delphi study produced consensus principles for responsible student use (Simoni et al., 2026), and a structured AI ethics programme improved nursing students' ethical competencies (Abuadas et al., 2025). A set of principles and a single quasi-experiment do not yet show that accountable behaviour carries over into clinical practice, but the two studies indicate that accountability can be taught rather than left to chance.

2.3.4 Cognitive offloading and dependency

Offloading has now been examined empirically, although mostly through self-report. In a three-wave longitudinal study of nursing students, rising GenAI dependency predicted lower metacognitive awareness, which carried through to self-perceived clinical decision-making (Xin et al., 2027). In Ibadan, daily AI users were more likely than occasional users to accept AI answers without verification (62.8% vs. 41.8%), although AI-use frequency was also modestly associated with higher cognitive reflection (Damilola & Aisagbonhi, 2026). In a single-centre survey, a verification habit buffered familiarity-driven overdependence (Popa et al., 2026). Outside education, routine AI exposure in colonoscopy was followed by a fall in adenoma detection during non-AI procedures from 28.4% to 22.4% (Budzyń et al., 2025). That study concerns experienced clinicians, not learners, but it shows how quickly an established skill can erode once a tool is routinely available; for learners, the greater risk is that the skill never forms. Counter-evidence should also be noted: designed GenAI use improved nursing students' higher-order thinking (Song et al., 2025), and a scoping review concludes that GenAI plays a dual role, with effects that depend on how it is integrated (Ning et al., 2026).

2.3.5 GenAI and assessment

AI now affects assessment in two ways: it can answer examination items, and it can also write and grade such items. The second use receives less attention, but it means that the verification problem applies to faculty as well as to students. ChatGPT-generated questions had more non-functional distractors than faculty items (Chauhan et al., 2025). GPT-4o scored short-answer clinical reasoning responses equivalently to faculty overall but diverged on "understanding" and "evaluating" items (Rajan et al., 2026). Educators have proposed staged access, comparative human–AI tasks and "AI-free oral or in-class assessment" (Bruijn & Vlachopoulos, 2026), referred to in this paper as AI-off assessment.

2.3.6 Evidence from African settings

Twenty-one of the studies reviewed came from African settings or African-led teams. These studies showed high use with little formal instruction: 78.6% of nursing and midwifery students in Ghana used AI, mostly learning about it informally, and the authors identified a "usage–knowledge disconnect" (Johnson et al., 2026). In a South African teaching-hospital survey, only 20 of 360 health professionals involved in teaching reported prior ChatGPT training, and 64.5% were concerned about reduced independent thinking (Maharaj et al., 2026). A multi-university Ugandan survey linked ChatGPT use and reported verification to higher self-rated clinical reasoning;

© IIARJournal. All rights reserved.

objective performance was not measured (Afam & Uzorka, 2026). The South African bedside trial is the only African randomised trial of clinical performance in this evidence base (Saloojee et al., 2026). The African evidence is therefore rich in surveys and thin in outcome studies: it documents use, attitudes and readiness, but apart from that trial it cannot yet show what GenAI does to learning in African classrooms and wards.

2.4 GAP IN THE LITERATURE

Taken together, the literature shows that unstructured GenAI access has not consistently improved learners' reasoning, that learners detect AI error poorly, and that dependency is associated with weaker metacognition. What it lacks is a framework that connects these findings to decisions about curriculum and assessment. Such a framework would need to treat reasoning, verification and accountability as linked competencies across medicine and nursing, to state for each competency whether it may be practised with AI and whether it must be certified without it, and to fit the resource and policy conditions of African schools. This paper sets out to meet these three requirements.

3. MATERIALS AND METHODS

3.1 RESEARCH DESIGN

This is a theoretical and conceptual study. It uses an integrative review of the 2023–2026 literature as its evidence base and builds a competency framework by theory-informed synthesis. It involved no human participants, intervention or primary data collection, so ethical approval was not required.

3.2 POPULATION AND SAMPLE

The population was peer-reviewed journal literature on GenAI in medical, nursing and health professions education published from 2023 to 2026. Screening ran in three rounds across the databases listed in Section 3.4. Round 1 covered approximately 390 unique MEDLINE titles, drawn from the 100 most-cited records for each domain string and from targeted searches. Round 2 covered 278 records not already retrieved (56 from ERIC and 222 from OpenAlex). Round 3 covered 225 PubMed records, approximately 40 Google Scholar results from the first results page of four queries, and the 100 most recent records from the Directory of Open Access Journals (DOAJ), all not already retrieved. The round 1 and Google Scholar counts are approximate because duplicates were removed during screening. Of 153 shortlisted articles, 10 were excluded (Section 3.3), leaving a final sample of 143: 139 in Scopus-indexed journals and four in peer-reviewed journals with Crossref-registered DOIs that are not indexed in Scopus. Fifteen were randomised trials or randomised crossover studies, 38 involved nursing or nursing-inclusive populations, and 21 came from African settings or African-led teams.

3.3 INCLUSION AND EXCLUSION CRITERIA

Eligible articles were peer-reviewed English-language journal articles that addressed at least one framework domain in medical, nursing or health professions education, or addressed AI-induced deskilling or automation bias among clinicians. Priority was given to randomised and quasi-experimental studies, systematic and scoping reviews, nursing studies, competency or governance papers, and studies from African and other lower-resource settings. Letters, editorials, comments, news items and preprints were excluded, as were articles whose abstract or full text could not be read to confirm the reported findings (seven articles) and brief commentaries or general reviews without domain-specific findings (three articles). Titles and abstracts were screened by a single reviewer with GenAI support (Section 3.8); screening was not duplicated independently, but every shortlisted record passed the verification checks described in Section 3.6.

3.4 DATA COLLECTION APPROACH AND TOOLS

Searches were run on 20 October 2026 in Europe PMC (MEDLINE), PubMed with Medical Subject Headings (MeSH), ERIC, OpenAlex, Google Scholar and DOAJ. Search strings combined a GenAI block with a medical, nursing and health professions education block. Each combination was then paired with one of five domain blocks: clinical reasoning, verification (accuracy, hallucination, reliability), accountability (ethics, integrity, professionalism), cognitive offloading (deskilling, automation bias, over-reliance, critical thinking) and assessment and competency (AI literacy, curricula,

© IIARJournal. All rights reserved.

frameworks). The offloading block was also run without the education block to capture clinical deskilling studies. Each included article was charted in a literature matrix recording setting, design, population, findings, relevance and verification status; full strings, hit counts and the matrix are given in Supplementary File 1.

3.5 DATA ANALYSIS AND BASES OF JUDGEMENT

Synthesis moved from the search domains to the competencies. Findings were first grouped deductively by the five search domains and then coded inductively for what the learner must be able to do, whether AI assistance showed benefit, harm or no effect, and what moderated the effect. The coded statements were consolidated into competencies and assigned to one of three clusters. The complementarity condition served as the basis of judgement: a competency was classed as one to be certified unaided when the evidence indicated that learners could not evaluate AI output without that competency. The resulting framework was then compared with five published frameworks to identify areas of overlap and of distinct contribution.

3.6 RELIABILITY AND VALIDITY

Every article passed five checks before inclusion. Its DOI was resolved on the Crossref REST API, and the authors, title, journal, volume, issue and pagination were taken from the publisher record rather than from secondary databases. The journal's Scopus status was checked in the SCImago Journal & Country Rank. Crossref update metadata, which incorporates Retraction Watch data, was searched for retractions, corrections and expressions of concern, and the first and last authors were searched in Europe PMC for retraction patterns. Each record was then confirmed to carry no retraction notice. No included article was retracted, withdrawn or subject to an expression of concern. Three carry a published erratum or correction, which was checked before citation (Budzyń et al., 2025; Singla et al., 2024; Zack et al., 2024). Three foundational theoretical sources published before 2023 (Miller, 1990; Parasuraman & Manzey, 2010; Risko & Gilbert, 2016) were cited for theory only and verified by the same procedure. Every cited finding was then checked against the original article abstract and, where available, the full text. Validity is limited because subscription databases, including Scopus, Web of Science and CINAHL, could not be searched directly, although every included journal was checked against Scopus.

3.7 VARIABLE MEASUREMENT

As a conceptual study, the paper measures no variables. Its constructs (GenAI, clinical reasoning, verification, accountability and cognitive offloading) are defined in Section 2.1, and the 12 competencies derived from these constructs are specified as observable learner behaviours with assessment modes in Table 2.

3.8 AI USAGE DISCLOSURE

In line with the framework's own accountability principle, the use of GenAI is disclosed. Claude (Anthropic), a GenAI assistant, was used to support database searching, metadata retrieval from Crossref, screening and drafting of the text. No AI tool is listed as an author. The author(s) checked all sources against the publisher records, verified every cited finding against the original sources and take full responsibility for the content.

4. RESULTS AND DISCUSSION

4.1 RESULTS

4.1.1 Evidence synthesis

Table 1 summarises the evidence for each domain and the implication drawn for the framework. The common thread is that GenAI was beneficial when the task sustained the learner's own reasoning, and of little benefit or even harmful when it did not. Verification was weak but appeared teachable, and accountability norms were being acquired largely through informal learning.

© IIARJournal. All rights reserved.

Table 1. Evidence synthesis by framework domain (143 verified articles, 2023–2026)

Domain	*n*	Principal evidence	Implication for the framework
Clinical reasoning	27	Unstructured LLM access gave no benefit in three trials (Goh et al., 2024; Gustafsson et al., 2026; Saloojee et al., 2026); structured protocols, simulation and feedback improved reasoning (Brügge et al., 2024; Lai et al., 2026; Z. Wang et al., 2025); pooled RCTs: knowledge SMD 0.99 (low certainty), clinical skills SMD 1.09 (moderate), follow-up SMD 0.51 (C. Wang et al., 2026).	Learners must reason first and use AI inside designed activities; independent reasoning is certified unaided.
Verification	20	Hallucination detection near chance at 55.0% (Zhou et al., 2025); accuracy falls on multi-step and image items (Bodke & Jumle, 2026); a 10-minute tutorial improved LLM-assisted diagnosis (Kim et al., 2025).	Verification is weak but responds to brief instruction in small studies; error detection is certified, and checking processes are assessed.
Accountability	27	Uncertainty at integrity boundaries (Dodson et al., 2025); ethics learned informally (Muhalil et al., 2026); consensus principles and teachable ethics (Abuadas et al., 2025; Simoni et al., 2026).	Ownership, disclosure and integrity must be taught and assessed explicitly, supported by institutional policy.
Cognitive offloading	26	Dependency predicted lower metacognition over three waves (Xin et al., 2027); daily users verified less (Damilola & Aisagbonhi, 2026); designed use improved higher-order thinking (Song et al., 2025).	Offloading risk is real but design-dependent; AI-off practice and checkpoints protect skill acquisition.
Assessment and competency	43	No consensus on AI competencies (Simoni et al., 2025); AI-written items weaker than faculty items (Chauhan et al., 2025); competing frameworks for single professions or levels (Corral-Gudino et al., 2026; Xie et al., 2026).	A cross-professional framework with an explicit assisted–unaided assessment rule is needed.

Note. *n* = number of articles in the literature matrix assigned to each domain. LLM = large language model; RCTs = randomised controlled trials; SMD = standardised mean difference.

4.1.2 The Reason–Verify–Account (RVA) Competency Framework

The framework (Figure 1, Table 2) organises 12 competencies into three linked clusters. Each cluster protects part of the learner's capacity to evaluate AI output, which the complementarity condition (Section 2.2) requires.

Cluster R: Clinical reasoning that holds without the tool (R1–R4). R1 (forming an independent differential diagnosis first) and R2 (reasoning under uncertainty) are placed at the stages where LLMs are weakest (Rao et al., 2026) and where upskilling inhibition threatens the development of clinical reasoning (Turney et al., 2026). R3 adopts the ABCD protocol because it has been tested in a randomised trial (Lai et al., 2026), and R4 draws on the simulation and feedback trials (Brügge et al., 2024; Z. Wang et al., 2025).

Cluster V: Verification of machine output (V1–V4). V1 (checking AI output against authoritative evidence and guidelines) applies the "retrieve, summarise and verify" paradigm, which leaves the user responsible for checking LLM output (Jin et al., 2023). V2 (detecting characteristic AI errors) answers the near-chance detection of hallucinations (Zhou et al., 2025) and the documented patterns of AI error (Saban & Dubovi, 2025). V3 (fit to patient, language and context, including bias) follows from evidence that performance depends on language and context (Cherif et al., 2024) and that models can reproduce bias (Zack et al., 2024). V4 (self-monitoring of reliance) turns the protective verification habit observed by Popa et al. (2026) into an explicit learning outcome.

Cluster A: Accountability for decisions the learner did not generate (A1–A4). A1 (owning and defending the decision) and A2 (disclosure) adopt the "disclose, verify, and revise under human authorship" model (Ali & Shaban, 2026) and the layered accountability proposed by Xu et al. (2026). A3 (privacy, consent and fairness) aligns with consensus ethical principles for student use (Simoni et al., 2026), while A4 (integrity in boundary cases) targets the uncertainty that learners reported about where acceptable use ends (Dodson et al., 2025).

© IIARJournal. All rights reserved.

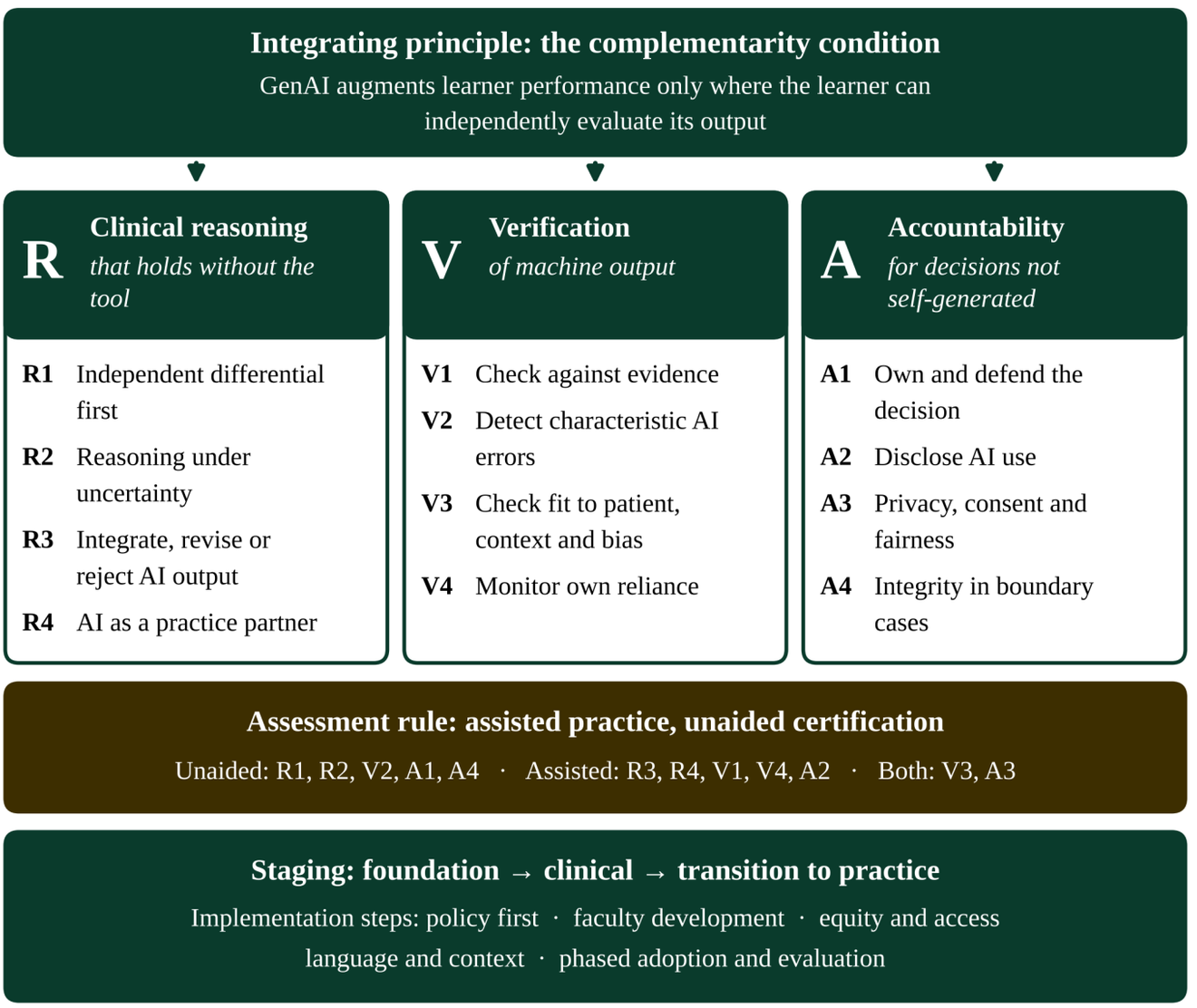


Figure 1. The Reason–Verify–Account (RVA) Competency Framework

Table 2. The Reason–Verify–Account (RVA) Competency Framework

Code	Competency: the learner can …	Assessment mode	Phase	Example assessment
R1	Form an independent problem representation and differential diagnosis before consulting AI	Unaided	Foundation	Key-features items; bedside or oral case presentation
R2	Reason under uncertainty and calibrate confidence, including when not to close a diagnosis	Unaided	Transition	Oral examination; script concordance test
R3	Integrate AI output by comparing, revising or rejecting it with stated reasons	Assisted	Clinical	Annotated AI transcript within a case write-up
R4	Use AI as a practice partner (simulated patients, feedback) within designed activities	Assisted	Foundation–clinical	Formative simulation log with reflection
V1	Check AI output against authoritative evidence and guidelines	Assisted	Foundation	Verification note citing primary sources
V2	Detect characteristic AI errors: hallucination, omission, inconsistency, over-triage	Unaided	Clinical	Station with errors planted in an AI-generated answer
V3	Check fit to the patient, language, local guidelines and resources, and screen for bias	Both	Clinical	Case adaptation task; OSCE station

© IIARJournal. All rights reserved.

Code	Competency: the learner can …	Assessment mode	Phase	Example assessment
V4	Monitor own reliance using repeat-and-compare and multi-source validation	Assisted	Transition	Reliance self-audit; validated reliance scale
A1	Own every signed AI-informed decision and defend it without the tool	Unaided	Transition	Viva on a submitted plan or note
A2	Disclose AI use in academic and clinical documentation	Assisted	Foundation	Disclosure statement checked against the transcript
A3	Use AI with regard for patient privacy, consent and fairness	Both	Clinical	Ethics case discussion; data-handling audit
A4	Apply integrity rules correctly, including in boundary cases	Unaided	Foundation	Short written boundary-case judgements

Note. Unaided = the learner may not use AI to perform the task, although AI-generated material may be its object. Assisted = the learner uses AI and is assessed on the process. Phase = the phase in which the competency is first certified: foundation (pre-clinical), clinical, or transition to practice. OSCE = objective structured clinical examination.

4.1.3 The assisted–unaided assessment rule

The framework's central assessment rule is *assisted practice, unaided certification* for the competencies that make evaluation of AI output possible. In this rule, "unaided" means that the learner may not use AI to perform the task, although AI-generated material may be the object of the task. Five competencies (R1, R2, V2, A1 and A4) must be certified unaided, because a learner unable to perform these tasks without assistance cannot meet the complementarity condition. These competencies are assessed through formats that already exist and cost little: bedside and oral examinations, key-features and script concordance items, objective structured clinical examinations (OSCEs), and, for V2, stations in which the learner must find errors planted in an AI-generated answer. Planted-error tasks draw on the finding that exposure to AI flaws builds critical appraisal (Çiçek et al., 2025). The AI-off formats reflect the checkpoints of Corral-Gudino et al. (2026) and the AI-off oral and in-class assessment proposed by educators (Bruijn & Vlachopoulos, 2026).

Five competencies (R3, R4, V1, V4 and A2) are assessed *with* AI, because the competence lies in how the tool is used. For these, the assessed evidence is the learner's process, such as prompt-and-response transcripts annotated with verification notes and accompanied by a disclosure statement. The remaining two (V3 and A3) are assessed in both modes. The rule is deliberately conservative. The rule may appear inconsistent with a future in which clinicians routinely work with AI; however, unaided certification does not imply unaided practice. Unaided certification provides the only means of establishing that the learner could recognise when the tool has failed. Faculty are bound by the same rule: AI-generated items and grades require human review before use (Chauhan et al., 2025; Rajan et al., 2026).

4.1.4 Developmental sequencing

The competencies are staged across three phases (Table 2). In the foundation (pre-clinical) phase, AI-off reasoning (R1), supervised practice with AI simulations (R4), evidence checking (V1), disclosure (A2) and integrity (A4) are prioritised, because learners at this stage are most exposed to never-skilling (Y. Liu et al., 2026). In the clinical phase, learners integrate and verify AI output against real patients (R3, V2, V3, A3). In the transition-to-practice phase, calibrated reasoning, self-monitored reliance and ownership (R2, V4, A1) are certified as conditions of entrustment, consistent with the stakes-by-autonomy approach of Patil et al. (2026).

4.1.5 Implementation path for African medical and nursing schools

The framework is designed to be implemented without new infrastructure, since its core certification formats, such as oral and bedside examination, require no AI tools or connectivity. Five implementation steps follow from the evidence reviewed in Section 2.3, and the sequence is deliberate.

Policy first. A school should publish a GenAI policy that states the assisted–unaided rule for each assessment before teaching practices are changed, because students currently acquire ethical norms informally (Muhalil et al., 2026) and have called for clear policies and modified assessment (Nantale et al., 2025). *Faculty development* should then precede student training. Only 20 of 360 surveyed South African health professionals reported prior ChatGPT training (Maharaj et al., 2026), and credible

© IIARJournal. All rights reserved.

assessment of verification requires educators who have practised the skill.

Equity and access should be treated as a design requirement rather than an afterthought, since unequal digital infrastructure and socio-economic access shape uptake (Ahmed et al., 2025). The same reasoning applies to *language and context*: verification should be taught against local guidelines and in the language of practice, given the difficulties reported with non-English assessment items (Cherif et al., 2024).

Phased adoption and evaluation allows schools to begin without waiting for ideal conditions (Adefolalu, 2026). Reliance can be measured with a validated instrument such as the Reliance on ChatGPT Scale for nursing students (Elzeky & Shahine, 2026), and evaluation should track learning outcomes, not user satisfaction alone.

4.2 DISCUSSION

4.2.1 Interpretation against the evidence and theory

The framework's main claim, that GenAI makes reasoning, verification and accountability more rather than less important, is supported by the trial evidence, though not without exception. The null and negative trials share unstructured access (Goh et al., 2024; Gustafsson et al., 2026; Saloojee et al., 2026), while most positive trials share structures that keep the learner reasoning (Arkan et al., 2025; Brügge et al., 2024; Lai et al., 2026). The rheumatology trial (Roemer et al., 2026) shows that the division is not absolute, so the framework is best read as a claim about where benefit is most reliable rather than as a rule without exceptions. The overall pattern fits an offloading account in which offloading helps when the user can evaluate the aid and harms when it displaces the processing that builds the skill (cf. Risko & Gilbert, 2016).

The verification findings accord with automation-bias theory (Parasuraman & Manzey, 2010). Trainees detected hallucinations near chance (Zhou et al., 2025), novices were most swayed by incorrect AI suggestions (Dratsch et al., 2023), and daily users verified less (Damilola & Aisagbonhi, 2026). The longitudinal link between dependency and lower metacognition (Xin et al., 2027) points to a mechanism that may connect the two theories: reduced self-monitoring of reasoning may be accompanied by reduced monitoring of machine output. If so, verification and metacognition are not separate educational goals, and teaching one should strengthen the other.

The evidence does not, however, justify a purely defensive stance. Designed use improved nursing students' higher-order thinking (Song et al., 2025), which suggests that GenAI can strengthen cognition as well as weaken it. The framework therefore places GenAI inside practice (R3, R4, V1) while protecting the competencies that make such practice safe.

4.2.2 Comparison with existing frameworks

Table 3 compares the RVA framework with five published frameworks. Its domains overlap substantially with the nursing framework of Xie et al. (2026), which already includes clinical reasoning, evidence verification, documentation integrity and ethical accountability, and with the AI-on and AI-off cycles of Corral-Gudino et al. (2026). Its contribution therefore lies less in new domains than in the way those domains are connected and applied. The RVA framework brings medicine and nursing into one structure, and it derives its assessment rule from a stated principle rather than from expert preference alone, so that the reason a competency must be certified unaided can be examined and contested. The framework also sequences the competencies across the programme and sets out an implementation path for African and other lower-resource schools, two matters that receive less attention in the comparator frameworks.

Table 3. Comparison of the RVA framework with published GenAI frameworks in health professions education

Framework	Target group	Structure	Relation to the RVA framework
Cao et al. (2026)	Clinicians in practice	Three tiers: safe use, evaluative expertise, governance and accountability leadership	Practice-level; RVA addresses pre-registration training
Kabbash et al. (2026)	Graduate medical education	AI content and assessment mapped across the six ACGME core competencies	Postgraduate; RVA covers undergraduate medicine and nursing

© IIARJournal. All rights reserved.

Framework	Target group	Structure	Relation to the RVA framework
Meyers & Miller (2026)	Medical education faculty	Six faculty GenAI competency domains	Faculty-side complement to RVA
Xie et al. (2026)	Nursing education	GANC framework: seven domains, including safety-critical clinical reasoning, evidence verification, documentation integrity and ethical accountability; simulation and OSCE; manual-mode readiness	Closest in content; RVA extends it to medicine and assigns an assessment mode to each competency
Corral-Gudino et al. (2026)	Medical education	Staged scaffold with AI-on and AI-off cycles and AI-off checkpoints	RVA adopts AI-off checkpoints and links these checkpoints to named competencies
RVA framework (this paper)	Medical and nursing education	12 competencies in three clusters, staged across three phases	Assisted–unaided rule per competency; implementation path for African schools

Note. ACGME = Accreditation Council for Graduate Medical Education; OSCE = objective structured clinical examination. Descriptions are summarised from the cited sources.

4.2.3 Propositions for testing

Because the framework is conceptual, its value depends on whether its claims withstand empirical testing. Four propositions follow. P1: Structured protocols that require independent reasoning before AI consultation will improve learners' decision quality, whereas unstructured access will not. P2: Verification competence will moderate the association between frequency of GenAI use and unverified acceptance or dependency. P3: Programmes that certify core reasoning unaided will show smaller deskilling and never-skilling effects than programmes that certify only AI-assisted performance. P4: Explicit accountability teaching will improve learners' judgement in boundary cases more than policy statements alone. P1 and P2 can be tested with existing designs and instruments (Elzeky & Shahine, 2026; Lai et al., 2026). P3 is harder to test, because deskilling and never-skilling unfold over years, and it will require longitudinal follow-up of the kind recommended by Turney et al. (2026).

4.2.4 Limitations

The framework is conceptual. Its competencies are derived from evidence, much of it cross-sectional or short-term, and have not yet been validated by expert consensus. Several outcome measures in the source studies were self-reported. The evidence base contains only one African randomised trial of clinical performance, and subscription databases could not be searched directly. The search was limited to English-language journal articles, screening was performed by a single reviewer, and some database runs were restricted to the most-cited or most recent records, so relevant studies may have been missed; publication bias towards positive findings is also possible. The field is changing quickly, and newer models may perform differently from those studied. The framework is intended to remain valid as tools change, because its competencies describe what learners must be able to do rather than how a particular tool behaves. Formal validation through a Delphi process of the kind used for curricular AI competencies (Singla et al., 2024), followed by testing of the propositions, is the necessary next step.

5. CONCLUSION AND RECOMMENDATION

GenAI can now answer many of the questions on which medical and nursing education has relied to certify competence. The evidence reviewed shows that this does not, by itself, make learners better clinicians. Unstructured access has not consistently improved reasoning, learners detect AI error poorly, and dependency is linked to weaker metacognition; benefit appears most reliably when learners continue to reason independently and check what the tool produces. The paper concludes that GenAI raises the importance of three capacities: clinical reasoning that holds without the tool, verification of machine output, and accountability for decisions the learner did not generate. The Reason–Verify–Account framework turns these into 12 competencies, specifies which must be certified unaided, sequences these competencies across training and sets out an implementation path for African schools.

The most immediate implications concern assessment. Medical and nursing schools should adopt the rule of assisted practice and unaided certification and state, for each assessment, whether AI is permitted. Verification deserves explicit and early teaching through exercises with planted AI

© IIARJournal. All rights reserved.

errors, since the skill is weak and early evidence suggests it responds to brief instruction. Institutions should also publish GenAI policies that require disclosure and human authorship, and should train faculty before students. Beyond individual schools, medical and nursing regulatory councils, together with researchers, should validate the framework through expert consensus and test its propositions in African settings using performance-based, longitudinal outcomes.

6. CONTRIBUTION TO EXISTING KNOWLEDGE

The paper's main theoretical contribution is the complementarity condition. The condition explains why structured and unstructured AI use have produced different results in trials and provides a principled basis for deciding what must be certified unaided. Building on this condition, the paper offers a single framework for medicine and nursing that links reasoning, verification and accountability as separately assessable competencies. For educators, the assisted–unaided rule can be applied with low-cost formats that most schools already use, and the implementation path and four testable propositions give African schools and researchers a concrete starting point.

7. REFERENCES

Abuadas, M., Albikawi, Z., & Rayani, A. (2025). The impact of an AI-focused ethics education program on nursing students' ethical awareness, moral sensitivity, attitudes, and generative AI adoption intention: A quasi-experimental study. *BMC Nursing*, *24*(1), Article 720. https://doi.org/10.1186/s12912-025-03458-2

Adefolalu, A. O. (2026). Integrating artificial intelligence into medical education in LMICs: A narrative review. *Advances in Medical Education and Practice*, *17*, 1–15. https://doi.org/10.2147/amep.s613617

Afam, G., & Uzorka, A. (2026). ChatGPT use, perceived usefulness, and self-reported clinical reasoning among undergraduate nursing students in Uganda: A cross-sectional study. *BMC Medical Education*. Advance online publication. https://doi.org/10.1186/s12909-026-10388-3

Ahmed, W. M. M., Abdalmotalib, M. M., Elbadawi, M. H., Mohammed, G. T. F., Mohamed, W. M. I., Mohammed, F. S. B., Salih, H. S., & Mohamed, H. O. Y. (2025). Shaping the future of medical education: A cross-sectional study on ChatGPT attitude and usage among medical students in Sudan. *PLOS One*, *20*(5), Article e0324698. https://doi.org/10.1371/journal.pone.0324698

Ali, S. I., & Shaban, M. (2026). Constructing digital professionalism in the generative AI era: A qualitative study with nursing students and faculty. *Nurse Education Today*, *162*, Article 107055. https://doi.org/10.1016/j.nedt.2026.107055

Arkan, B., Dallı, Ö. E., & Varol, B. (2025). The impact of ChatGPT training in the nursing process on nursing students' problem-solving skills, attitudes towards artificial intelligence, competency, and satisfaction levels: Single-blind randomized controlled study. *Nurse Education Today*, *152*, Article 106765. https://doi.org/10.1016/j.nedt.2025.106765

Bodke, P., & Jumle, R. (2026). Large language models in undergraduate medical education: A systematic review of examination performance, knowledge acquisition, and learner perceptions. *BMC Medical Education*. Advance online publication. https://doi.org/10.1186/s12909-026-09999-7

Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., Holling, M., Lu, M. H., & Darici, D. (2024). Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial. *BMC Medical Education*, *24*(1), Article 1391. https://doi.org/10.1186/s12909-024-06399-7

Bruijn, K., & Vlachopoulos, D. (2026). Between co-intelligence and cognitive offloading: Educators' perspectives from Dutch academic medical settings on generative AI, critical thinking, and adult professional learning. *Education Sciences*, *16*(9), Article 1470. https://doi.org/10.3390/educsci16091470

© IIARJournal. All rights reserved.

Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., . . . Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, *10*(10), 896–903. https://doi.org/10.1016/s2468-1253(25)00133-5

Cao, W., Zhang, Q., Liu, J., & Liu, S. (2026). From agents to governance: Essential AI skills for clinicians in the large language model era. *Journal of Medical Internet Research*, *28*, Article e86550. https://doi.org/10.2196/86550

Chauhan, A., Khaliq, F., & Nayak, K. R. (2025). Assessing quality of scenario-based multiple-choice questions in physiology: Faculty-generated vs. ChatGPT-generated questions among Phase I medical students. *International Journal of Artificial Intelligence in Education*, *35*(4), 2315–2344. https://doi.org/10.1007/s40593-025-00471-z

Cherif, H., Moussa, C., Missaoui, A. M., Salouage, I., Mokaddem, S., & Dhahri, B. (2024). Appraisal of ChatGPT's aptitude for medical education: Comparative analysis with third-year medical students in a pulmonology examination. *JMIR Medical Education*, *10*, Article e52818. https://doi.org/10.2196/52818

Çiçek, F. E., Ülker, M., Özer, M., & Kıyak, Y. S. (2025). ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: A randomized controlled trial. *Postgraduate Medical Journal*, *101*(1195), 458–463. https://doi.org/10.1093/postmj/qgae170

Corral-Gudino, L., Herrero-Montano, I., de la Torre-Díez, I., & Miramontes-González, J. P. (2026). Integrating generative artificial intelligence (AI) in medical education: A framework for preserving clinical reasoning. *Applied Sciences*, *16*(12), Article 5946. https://doi.org/10.3390/app16125946

Damilola, I. M., & Aisagbonhi, H. O. (2026). Artificial intelligence use and its cognitive and psychological correlates among medical and allied health students in Ibadan, Nigeria. *BMC Medical Education*. Advance online publication. https://doi.org/10.1186/s12909-026-10558-3

Dodson, T. M., Thompson-Hairston, K., & Reed, J. M. (2025). Nursing students' AI literacy and ethical understanding of AI in nursing education. *Teaching and Learning in Nursing*, *20*(4), 390–394. https://doi.org/10.1016/j.teln.2025.07.004

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, *307*(4), Article e222176. https://doi.org/10.1148/radiol.222176

El-Banna, M. M., Sajid, M. R., Rizvi, M. R., Sami, W., & McNelis, A. M. (2025). AI literacy and competency in nursing education: Preparing students and faculty members for an AI-enabled future—A systematic review and meta-analysis. *Frontiers in Medicine*, *12*, Article 1681784. https://doi.org/10.3389/fmed.2025.1681784

Elzeky, M. E. H., & Shahine, N. F. M. (2026). Reliance on ChatGPT in learning: Construction and validation of a reliance on ChatGPT scale for nursing students (ROCS-NS): Using structural equation modeling (SEM). *Nurse Education in Practice*, *91*, Article 104696. https://doi.org/10.1016/j.nepr.2025.104696

Feigerlova, E., Hani, H., & Hothersall-Davies, E. (2025). A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Medical Education*, *25*(1), Article 129. https://doi.org/10.1186/s12909-025-06719-5

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, *7*(10), Article e2440969.

© IIARJournal. All rights reserved.

https://doi.org/10.1001/jamanetworkopen.2024.40969

Gustafsson, J., Lehtonen-Smeds, E., & Pakkasjärvi, N. (2026). Integrating large language models into trauma education for medical students: Randomized controlled pilot trial. *JMIR Medical Education*, *12*, Article e79134. https://doi.org/10.2196/79134

Ichikawa, T., Olsen, E., Vinod, A., Glenn, N., Hanna, K., Lund, G. C., & Pierce-Talsma, S. (2025). Generative artificial intelligence in medical education—Policies and training at US osteopathic medical schools: Descriptive cross-sectional survey. *JMIR Medical Education*, *11*, Article e58766. https://doi.org/10.2196/58766

Iliyasu, Z., Abdullahi, H. O., Iliyasu, B. Z., Bashir, H. A., Amole, T. G., Abdullahi, H. M., Abdullahi, A. U., Kwaku, A. A., Dahir, T., Tsiga-Ahmed, F. I., Jibo, A. M., Salihu, H. M., & Aliyu, M. H. (2025). Correlates of medical and allied health students' engagement with generative AI in Nigeria. *Medical Science Educator*, *35*(1), 269–280. https://doi.org/10.1007/s40670-024-02181-y

Jin, Q., Leaman, R., & Lu, Z. (2023). Retrieve, summarize, and verify: How will ChatGPT affect information seeking from the medical literature? *Journal of the American Society of Nephrology*, *34*(8), 1302–1304. https://doi.org/10.1681/asn.0000000000000166

Johnson, B. B., Yeboah, J. O., Klutsey, E. E., Agbo, I., Lokpo, S. Y., Duneeh, R. V., Asem, L., Gyamfua, D. P. P., Kwadzokpui, P. K., Bohli, J. K., & Ablordey, K. (2026). Bridging the AI gap in nursing and midwifery education: A cross-sectional analysis of predictors of use and knowledge in Ghana. *Journal of Umm Al-Qura University for Medical Science*, *12*(2), Article 40. https://doi.org/10.1007/s44361-026-00053-1

Kabbash, M. S., Loftus, T. J., Morrow, J., Dent, D., & Rodriguez, R. (2026). Integrating artificial intelligence competencies across the six ACGME core competencies: A proposed framework for graduate medical education. *Advances in Medical Education and Practice*, *17*, 1–8. https://doi.org/10.2147/amep.s640400

Kim, S. H., Schramm, S., Wihl, J., Raffler, P., Tahedl, M., Canisius, J., Luiken, I., Endrös, L., Reischl, S., Marka, A., Walter, R., Schillmaier, M., Zimmer, C., Wiestler, B., & Hedderich, D. M. (2025). Boosting LLM-assisted diagnosis: 10-minute LLM tutorial elevates radiology residents' performance in brain MRI interpretation. *Neuroradiology*, *67*(8), 2069–2081. https://doi.org/10.1007/s00234-025-03664-4

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), Article e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lai, S. Y., Fan, Y., Yao, F., & Zhang, S. (2026). The ABCD of AI-enabled clinical reasoning: A guided framework to transform learner interactions and decision quality. *Journal of Surgical Education*, *83*(6), Article 103963. https://doi.org/10.1016/j.jsurg.2026.103963

Li, X., Yan, X., & Lai, H. (2025). The ethical challenges in the integration of artificial intelligence and large language models in medical education: A scoping review. *PLOS One*, *20*(10), Article e0333411. https://doi.org/10.1371/journal.pone.0333411

Liu, M., Okuhara, T., Chang, X., Shirabe, R., Nishiie, Y., Okada, H., & Kiuchi, T. (2024). Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *26*, Article e60807. https://doi.org/10.2196/60807

Liu, Y., Han, L., Che, L., Dong, W., Zhang, T., & Guo, H. (2026). Cognitive reshaping and resurgence of humanness: Restructuring the medical education continuum in the era of generative AI. *Frontiers in Medicine*, *13*, Article 1867722. https://doi.org/10.3389/fmed.2026.1867722

Maharaj, N. R., Saaiman, C. M. B., & Hagemeister, D. T. (2026). Using Chat Generative Pre-

© IIARJournal. All rights reserved.

Trained Transformer for teaching and learning: A survey among health professions educators. *Journal of the Colleges of Medicine of South Africa*, *4*(1), Article a442. https://doi.org/10.4102/jcmsa.v4i1.442

Meyers, P., & Miller, C. (2026). Synthesising generative artificial intelligence competency domains for medical education faculty. *The Clinical Teacher*, *23*(6), Article e70532. https://doi.org/10.1111/tct.70532

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63–S67. https://doi.org/10.1097/00001888-199009000-00045

Muhalil, A. S. G., Alfaki, A. A. B. A., Ali, A. A. B., Alslawy, Y. M. A., Albashir, A. A. A., Abdelgadir, A. M. H., Ahmed, A. E. S., Abdalgader, H. A. S., Abdallah, R. M. A.-E., & Mohamed, Y. A. A. (2026). Artificial intelligence ethical awareness and academic integrity among medical students at the University of Khartoum: A cross-sectional study. *BMC Medical Education*. Advance online publication. https://doi.org/10.1186/s12909-026-10617-9

Nantale, R., Kayemba, F., Paungholi, K., Epuitai, J., Jemba, B., Ajalo, E., Afodun, A., Musaba, M. W., Munabi, I. G., Kiguli, S., Mubuuke, A. G., & Mukunya, D. (2025). Experiences of nursing and medical students with the use of artificial intelligence tools such as ChatGPT in Uganda. *PLOS Global Public Health*, *5*(12), Article e0005715. https://doi.org/10.1371/journal.pgph.0005715

Ning, Y., Sun, S., Lu, Y., Dong, W., Kong, C., & Liu, Y. (2026). Mapping the evidence on generative artificial intelligence for higher-order thinking skills in nursing students: A scoping review. *Nurse Education in Practice*, *96*, Article 104948. https://doi.org/10.1016/j.nepr.2026.104948

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Patil, S., Karabacak, M., Bisdas, S., & Margetis, K. (2026). Generative language models in medical education: From advent to entrustment. *JMIR Medical Education*, *12*, Article e105347. https://doi.org/10.2196/105347

Popa, D.-I., Levai, C. M., Buleu, F., Burtic, S., Militaru, M., Juganaru, I., & Bratu, M. L. (2026). Assessing generative AI adoption, tool preferences, and cognitive reliance among medical students: A cross-sectional study. *International Medical Education*, *5*(3), Article 66. https://doi.org/10.3390/ime5030066

Rajan, A., Alexander, S. M., & Shenvi, C. L. (2026). Can AI grade like a professor? Comparing artificial intelligence and faculty scoring of medical student short-answer clinical reasoning exams. *Advances in Health Sciences Education*, *31*(2), 607–617. https://doi.org/10.1007/s10459-025-10462-3

Rao, A. S., Esmail, K. P., Lee, R. S., Jiang, S., Arraiza Carlo, B., Gill, J., Khanna, P., Kalmowitz, E., Montagnese, B., Heydari, K., Jiao, Q., Bott, E., Nguyen, D., Wang, G., Hood, M., Landman, A. B., & Succi, M. D. (2026). Large language model performance and clinical reasoning tasks. *JAMA Network Open*, *9*(4), Article e264003. https://doi.org/10.1001/jamanetworkopen.2026.4003

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

Roemer, A., Schlicker, N., Kernder, A., Albe, B., Hack, J., Hirsch, M., Mayr, A., Kuhn, S., & Knitza, J. (2026). Large language models enhance diagnostic reasoning of medical students in rheumatology: A randomized controlled trial. *BMC Medical Education*, *26*(1), Article 579. https://doi.org/10.1186/s12909-026-09079-w

Saban, M., & Dubovi, I. (2025). A comparative vignette study: Evaluating the potential role of a generative AI model in enhancing clinical decision-making in nursing. *Journal of Advanced Nursing*, *81*(11), 7489–7499. https://doi.org/10.1111/jan.16101

© IIARJournal. All rights reserved.

Saloojee, H., Gramanie, M. C., Mwale, R., Bassett, B. A., Madhi, S. A., & Kalla, I. S. (2026). AI at the bedside: Randomised controlled trial of ChatGPT's impact on student performance in real-patient clinical exams. *Medical Teacher*. Advance online publication. https://doi.org/10.1080/0142159x.2026.2652061

Sengul, T., Sariköse, S., & Gul, A. (2025). Ethical decision-making and artificial intelligence in nursing education: An integrative review. *Nursing Ethics*, *32*(8), 2490–2515. https://doi.org/10.1177/09697330251366600

Simoni, J., Pereira, J. L., Aschenbrenner-Noriega, I., Crespo Sánchez, S., Egaña-Yin, D., García-Gargallo, C., Hertog, O. L. A., Idoate, A., Morales-Gutiérrez, I., Romero, A., Santiago-Martínez, P., Sawczyn, G., Vahamaki, A., Urtubia-Fernandez, J., Zurita, R., & Mengual, E. (2026). Principles for medical students' responsible use of generative AI: A student-partnered Delphi study. *BMC Medical Ethics*, *27*(1), Article 203. https://doi.org/10.1186/s12910-026-01550-z

Simoni, J., Urtubia-Fernandez, J., Mengual, E., Simoni, D. A., Royo, M., Egaña-Yin, D., Hertog, O. L. A., López-Ortiz, L., Muñoz-Tomás, A., Santiago-Martínez, P., Vahamaki, A., & Pereira, J. L. (2025). Artificial intelligence in undergraduate medical education: An updated scoping review. *BMC Medical Education*, *25*(1), Article 1609. https://doi.org/10.1186/s12909-025-08188-2

Singla, R., Pupic, N., Ghaffarizadeh, S.-A., Kim, C., Hu, R., Forster, B. B., & Hacihaliloglu, I. (2024). Developing a Canadian artificial intelligence medical curriculum using a Delphi study. *npj Digital Medicine*, *7*(1), Article 323. https://doi.org/10.1038/s41746-024-01307-1

Song, D., Zhang, P., Zhu, Y., Qi, S., Yang, Y., Gong, L., & Zhou, L. (2025). Effects of generative artificial intelligence on higher-order thinking skills and artificial intelligence literacy in nursing undergraduates: A quasi-experimental study. *Nurse Education in Practice*, *88*, Article 104549. https://doi.org/10.1016/j.nepr.2025.104549

Turney, J., Young, T., Chauhan, D., Beeharry, R., & Mahmud, M. (2026). AI use for medical students: Impact on clinical skill acquisition and retention. A systematic review. *Advances in Medical Education and Practice*, *17*, 1–13. https://doi.org/10.2147/amep.s583763

Wang, C., Sun, N., Pan, X., Liu, P., Ma, Z., Wang, W., & Jin, Y. (2026). The impact of integrating generative artificial intelligence into medical education on short-term learning outcomes: A systematic review and meta-analysis of randomized controlled trials. *BMC Medical Education*, *26*(1), Article 983. https://doi.org/10.1186/s12909-026-09320-6

Wang, Z., Fan, T.-T., Li, M.-L., Zhu, N.-J., & Wang, X.-C. (2025). Feasibility study of using GPT for history-taking training in medical education: A randomized clinical trial. *BMC Medical Education*, *25*(1), Article 1030. https://doi.org/10.1186/s12909-025-07614-9

Xie, W., Liu, F., Liu, J., & Liu, S. (2026). Beyond literacy to clinical competency: A framework for integrating generative AI into nursing education. *Nurse Education Today*, *163*, Article 107108. https://doi.org/10.1016/j.nedt.2026.107108

Xin, Y., Yan, D., Shuren, L., Weina, H., Shusheng, D., Minyang, L., & Liuheng, L. (2027). Generative AI dependency and self-perceived clinical decision-making in nursing students: A three-wave longitudinal study of metacognitive awareness as a mediator. *Nurse Education Today*, *168*, Article 107347. https://doi.org/10.1016/j.nedt.2026.107347

Xu, J., Zhang, M., Zheng, B., Deng, W., & Shi, R. (2026). Reclaiming human clinical reasoning in generative AI-enabled healthcare: A perspective on cognitive agency, skill evolution, and accountability. *Frontiers in Public Health*, *14*, Article 1979099. https://doi.org/10.3389/fpubh.2026.1979099

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation

© IIARJournal. All rights reserved.

study. *The Lancet Digital Health*, *6*(1), e12–e22. https://doi.org/10.1016/s2589-7500(23)00225-x

Zhou, J., Zhang, J., Wan, R., Cui, X., Liu, Q., Guo, H., Shi, X., Fu, B., Meng, J., Yue, B., Zhang, Y., & Zhang, Z. (2025). Integrating AI into clinical education: Evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. *BMC Medical Education*, *25*(1), Article 406. https://doi.org/10.1186/s12909-025-06916-2

© IIARJournal. All rights reserved.