American Journal of Medical Education
2026, Vol. 2, No. 3, pp. 50–60
https://doi.org/10.11648/j.mededu.20260203.12

Review Article

Teaching Health Professions Students When Not to Trust Generative AI: A Theory-Informed Framework for Verification-Scaffolded Case-Based Learning

Adesegun Nurudeen Osijirin[1, *], Shamsuddeen Mohammed Sada[2], Utibe Edmond Victor[3], Oka Obasi[4], Nsangaette Asuquo Udo[5]

[1]Department of Healthcare Management, Federal University of Allied Health Sciences, Enugu, Nigeria
[2]Department of Software Engineering, Federal University of Allied Health Sciences, Enugu, Nigeria
[3]Department of Health Information Management, Federal University of Allied Health Sciences, Enugu, Nigeria
[4]Department of Biomedical Engineering, Federal University of Allied Health Sciences, Enugu, Nigeria
[5]Department of Artificial Intelligence, Federal University of Allied Health Sciences, Enugu, Nigeria

Abstract

Background: Generative artificial intelligence (GenAI) can produce fluent, context-responsive explanations and clinical suggestions, but the same fluency can conceal factual, evidentiary, or reasoning errors. Health professions students therefore need more than access skills or generic warnings: they need repeated practice deciding when AI advice is warranted, when it requires verification, and when it should be rejected. *Aim:* To propose a theory-informed, operational framework for teaching appropriate reliance on GenAI through case-based learning. *Approach:* A focused critical synthesis integrated four bodies of scholarship: GenAI in health professions education, trust and automation bias in human-AI interaction, metacognition and self-regulated learning, and case-based learning. The purpose was conceptual integration rather than exhaustive evidence retrieval or estimation of intervention effects. *Framework:* Verification-Scaffolded Case-Based Learning (VS-CBL) is a six-phase cycle: Commit, Consult, Challenge, Corroborate, Calibrate, and Consolidate. Students first make and rate an unaided judgment, then consult a documented AI system. They actively test the response for omissions and counterevidence, verify pivotal claims against authoritative sources, decide whether to accept, modify, or reject the advice, and complete an accountable final response plus an unaided transfer task. Faculty prevalidate cases, vary AI accuracy and error subtlety, preserve patient privacy, and scale oversight to clinical risk. *Implications:* VS-CBL shifts the educational target from AI use or satisfaction to appropriate reliance. It supports measurement of overreliance, underreliance, verification quality, confidence-accuracy calibration, reasoning quality, and delayed independent transfer. The design is adaptable across medicine, nursing, pharmacy, dentistry, rehabilitation, public health, and allied health programmes, including settings with limited connectivity. *Conclusion:* The educational question is not whether students should trust GenAI in general, but whether trust is warranted for a particular claim, task, and level of risk. VS-CBL offers a testable curriculum architecture for making that judgment visible and teachable. Its effectiveness and implementation burden require prospective, multicentre evaluation.

*Correspondence: Adesegun Nurudeen Osijirin (adesegunosijirin@fuahse.edu.ng)

Received: 3 August 2026; **Accepted:** 14 August 2026; **Published:** 9 November 2026

Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Keywords

Generative Artificial Intelligence, Health Professions Education, Case-based Learning, Automation Bias, Metacognition, Trust Calibration, Clinical Reasoning, AI Literacy

1. Introduction

Generative artificial intelligence (GenAI), a form of artificial intelligence (AI), has moved from a peripheral technology to an everyday learning resource. Health professions students use large language models to explain concepts, generate cases and quizzes, compare diagnoses, rehearse communication, and obtain feedback. Reviews describe substantial educational promise, but they also show a young and uneven evidence base dominated by short-term outcomes, heterogeneous interventions, and limited attention to what students can do after the tool is removed [1-4]. The central educational risk is therefore not simply that GenAI can be wrong. It is that a plausible response can change a learner's judgment before its accuracy, evidentiary basis, or fit to the local clinical context has been established [5].

Recent experiments illustrate this distinction between access and effective use. In a randomized clinical trial, access to a large language model did not significantly improve physicians' diagnostic reasoning compared with conventional resources, even though the model performing alone achieved a high score [6]. In another randomized trial, ChatGPT-generated feedback produced overall learning outcomes similar to expert feedback for first-year medical students, but expert feedback was superior on more complex urinary tract infection cases [7]. Beyond health professions education, experimental evidence shows that unguarded GenAI can improve performance while assistance is available yet impair subsequent unaided performance; a safeguarded tutor design reduced that harm [8]. These findings do not justify either blanket enthusiasm or blanket prohibition. They point to a curriculum problem: learners must be taught how to regulate reliance.

Reliance is consequential in health professions education because students are still constructing the knowledge structures needed to recognise an error. Inaccurate decision support can degrade judgment, and people with less relevant expertise may be especially susceptible to plausible incorrect advice [9, 10]. Generic instructions to "check the answer" are weak safeguards when a learner does not know which claim is pivotal, which source is authoritative, or how much verification is proportionate to the stakes. Trust in automation is appropriately calibrated only when reliance tracks the system's capability and the demands of the task [11, 12]. In education, that calibration must coexist with a second goal: preserving the cognitive work through which durable expertise develops.

This article proposes Verification-Scaffolded Case-Based Learning (VS-CBL), a theory-informed framework that turns verification and reliance decisions into visible learning performances. The contribution is not another catalogue of GenAI opportunities. It is an operational sequence for teaching students when not to trust an AI response, when to verify it, and when reliance is warranted. The framework integrates cognitive forcing, metacognitive monitoring, appropriate-reliance theory, and guided case-based learning. It also supplies a case-design matrix, an assessment blueprint, implementation safeguards, and a research agenda suitable for international and resource-diverse health professions programmes.

2. Methods and Theoretical Foundations

2.1. Approach to Framework Development

VS-CBL was developed through a theory-informed critical synthesis rather than a systematic or scoping review. Targeted searches of PubMed and publisher platforms, supplemented by backward reference checking, were used to locate recent reviews and empirical studies of GenAI in health professions education and case-based learning. Seminal and contemporary work on automation bias, trust in automation, cognitive offloading, cognitive forcing, self-regulated learning, metacognition, and case-based learning was then purposively integrated. Searches were last updated on 3 August 2026. No study-count flow diagram or claim of exhaustive retrieval is presented because the purpose was to connect explanatory mechanisms to an implementable educational design, not to estimate an intervention effect.

Framework development followed three questions. First, what observable learner behaviours distinguish safe, educationally productive AI use from passive acceptance? Second, which instructional mechanisms can protect independent reasoning while allowing authentic AI use? Third, which outcomes would reveal appropriate reliance rather than mere improvement in an AI-assisted score? The proposed phases were iteratively aligned with those questions and checked against two constraints: each phase had to yield an assessable artefact, and the framework had to remain usable with different GenAI products and across health professions.

Ethics approval, consent to participate, and consent for publication were not applicable because this conceptual review

synthesised published literature and involved no human participants, patient data, or animals. ChatGPT (OpenAI; accessed 3 August 2026) supported structural development, drafting, and language editing; the authors critically revised the manuscript, verified scholarly claims and references, and take responsibility for the final content.

2.2. The Educational Problem: Performance Is Not the Same as Learning

GenAI can function as an external source of explanations, options, and language. That function may raise immediate task performance while reducing the learner's opportunity to retrieve knowledge, generate hypotheses, tolerate uncertainty, or notice gaps. Educational effectiveness cannot therefore be inferred from the quality of a jointly produced answer. A student may submit an excellent response by copying reliable assistance, a poor response by following unreliable assistance, or a superficially independent response that was anchored by an earlier AI suggestion. Each pathway has a different implication for learning and safety.

The relevant outcome is not maximal agreement with AI. It is appropriate reliance: accepting useful advice for the right reasons, rejecting misleading advice for the right reasons, and escalating uncertainty when available evidence is insufficient. Overreliance occurs when a learner accepts incorrect or unsuitable advice. Underreliance occurs when correct, useful advice is rejected. Appropriate reliance includes both accepting correct advice and resisting incorrect advice, with the important qualification that even technically correct advice may be inappropriate when it ignores patient values, local resources, scope-of-practice rules, or current guidance. Trust is thus relational and conditional—not a stable attribute of a model [11, 12].

2.3. Why Plausible AI Output Is Educationally Hazardous

Large language models generate probable linguistic sequences, not guaranteed facts or clinically accountable judgments. They may produce intrinsic errors that contradict supplied information, extrinsic claims that cannot be supported, fabricated references, outdated recommendations, or reasoning that is internally fluent but clinically incomplete [13-15]. A student can also receive a factually correct answer that is pedagogically harmful because it removes the need to construct a problem representation or compare alternatives. The framework therefore treats accuracy as necessary but insufficient. Educational safety also depends on timing, transparency, verification, task stakes, and whether learning transfers without assistance.

2.4. Mechanisms of Inappropriate Reliance

2.4.1. Automation Bias and Anchoring

Automation bias is the tendency to use automated advice as a heuristic substitute for vigilant information seeking and processing. It can appear as commission errors, in which incorrect advice is followed, or omission errors, in which a problem not flagged by the system is overlooked [10]. In case discussions, the first diagnosis or management plan offered by AI may become an anchor. Subsequent evidence is then interpreted around that suggestion, particularly when the response is fluent, detailed, or framed confidently. Asking for verification after exposure may be too late: the initial suggestion has already changed the search space.

VS-CBL addresses this timing problem by requiring a pre-AI commitment. Students create an unaided problem representation, identify discriminating evidence, state a provisional decision, and rate confidence before opening the model. This does not freeze judgment; revision is expected. It preserves an independent trace against which the direction and quality of later change can be assessed.

2.4.2. Cognitive Offloading and Fluency

Cognitive offloading is the use of external action or tools to reduce internal information processing demands [16]. Offloading is not inherently undesirable: clinicians appropriately use calculators, reference systems, and checklists. The concern is unregulated offloading of the very process being learned. Search technologies can shift what people remember and inflate estimates of internal knowledge [17, 18]. GenAI intensifies this possibility by presenting integrated explanations rather than a list of sources, reducing the visible effort between question and answer. The fluency of the result can be mistaken for the learner's own understanding.

VS-CBL introduces productive friction at points where learning would otherwise disappear. Students must reason before consultation, interrogate rather than merely receive the output, and later solve a related transfer case without AI. Friction is deliberately selective. It should protect high-value reasoning, not create busywork or force exhaustive checking of trivial claims.

2.4.3. The Novice Verification Paradox

The people who most need to verify an answer may have the least developed schema for detecting what is wrong. Novices can be impressed by technical vocabulary, lengthy differentials, or cited-looking text without knowing whether a source exists or a recommendation applies. This verification paradox means that responsibility cannot be shifted to the student through a disclaimer. Faculty must design cases in which errors are discoverable using resources that learners have been taught to evaluate. Early cases can use explicit verification prompts and a small approved evidence set; advanced cases can progressively remove prompts and require independent source selection.

2.4.4. Confidence Miscalibration

Confidence and accuracy are related but not interchangeable. High confidence in an incorrect answer is especially concerning, whereas low confidence in a correct answer may lead to unnecessary deference. Metacognitive calibration can be made visible by collecting confidence before and after AI exposure and comparing it with performance. An increase in confidence is not automatically beneficial: confidence that rises after incorrect advice represents harmful miscalibration. The framework therefore requires students to explain why confidence changed and which evidence warrants that change.

2.5. Theoretical Foundations of VS-CBL

2.5.1. Cognitive Forcing

Cognitive forcing functions introduce a constraint or deliberate step that interrupts an intuitive path. In human-AI decision tasks, forcing people to make an initial decision or engage analytic reasoning can reduce overreliance, although users may experience greater effort and prefer a less demanding interface [19]. Clinical reasoning scholarship similarly uses cognitive forcing and reflective strategies to counter predictable biases [20, 21]. The Commit and Challenge phases translate these ideas into educational actions: generate before viewing, then actively seek disconfirming information.

2.5.2. Self-regulated Learning and Metacognition

Self-regulated learning involves forethought, performance monitoring, and reflection [22, 23]. VS-CBL distributes these functions across the case. Commit externalises forethought; Consult and Challenge make strategy use visible; Corroborate requires monitoring against evidence; Calibrate requires judgment about confidence and reliance; and Consolidate supports reflection and transfer. Confidence-calibration tools can make discrepancies between perceived and demonstrated knowledge visible [24]. The cycle is therefore more than a checklist for fact-checking. It teaches students to monitor how an external system changes their thinking.

2.5.3. Appropriate Reliance and Trust Calibration

Trust should be calibrated to demonstrated system performance, task characteristics, user expertise, and consequences of failure [11, 12]. A model may be useful for brainstorming low-stakes communication variants yet unsuitable as the final authority for a time-critical drug dose. The same output may deserve different levels of scrutiny when used for private revision, a graded assignment, or patient care. VS-CBL therefore couples the accept-modify-reject decision with a risk gradient. As potential harm, irreversibility, uncertainty, or local variation increases, required corroboration and human oversight also increase.

2.5.4. Case-based Learning as the Instructional Carrier

Case-based learning is a guided inquiry approach in which learners apply knowledge to authentic or realistic problems. Reviews support its value for engagement, knowledge application, and clinical reasoning across health professions, while recognising heterogeneity in design and outcomes [25-27]. Cases are especially suitable for calibrated reliance because they contain contextual cues, incomplete information, competing explanations, and consequences. They allow the educator to manipulate AI accuracy and error subtlety without exposing patients to risk. Recent mapping of AI-supported case-based learning identifies growing use but also major gaps concerning automation bias, cognitive offloading, overreliance, and diagnostic calibration [4]. Reviews of AI education also find sparse use of learning theory and little evidence of workplace behaviour [28]. VS-CBL is designed to operationalise those unresolved outcomes.

3. Results

3.1. Verification-Scaffolded Case-Based Learning

VS-CBL is a six-phase cycle—Commit, Consult, Challenge, Corroborate, Calibrate, and Consolidate (Figure 1). The sequence is fixed at first exposure because timing is part of the intervention: independent judgment precedes AI advice. With increasing expertise, the amount of prompting can fade, but the core functions remain. Every phase produces evidence that can be discussed or assessed.

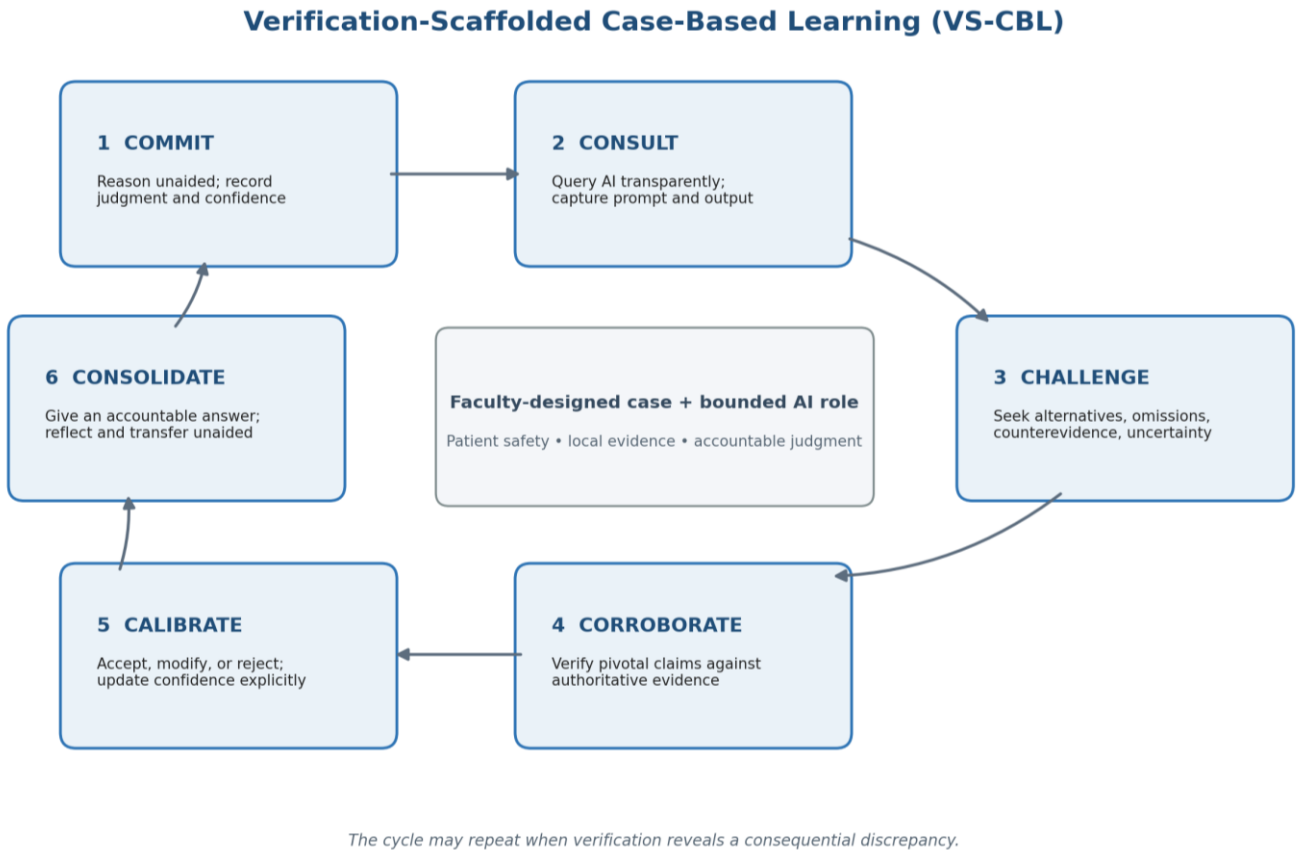


Figure 1. *The Verification-Scaffolded Case-Based Learning cycle. The cycle may repeat when corroboration reveals a consequential discrepancy. AI advice remains bounded by faculty design, patient safety, local evidence, and accountable human judgment. Original framework proposed in this article.*

Table 1. *Operational phases of Verification-Scaffolded Case-Based Learning.*

Phase	Instructional purpose	Student action	Educator role	Observable output
1. Commit	Protect independent generation and establish a baseline.	Write an unaided problem representation, options or differential, intended action, key evidence, and confidence (0-100%).	Withhold AI output; define the decision point and minimum reasoning artefact.	Timestamped initial judgment and confidence.
2. Consult	Use AI transparently as advice, not authority.	Submit a documented prompt; record product/model, date, prompt, output, and any retrieval mode.	Specify permissible data and tools; prevent entry of identifiable patient information.	Reproducible interaction record.
3. Challenge	Counter anchoring and surface uncertainty.	Ask for alternatives, missing information, counterevidence, assumptions, possible bias, and conditions under which advice would change.	Model adversarial questioning; include plausible partial and subtle errors.	Challenge log identifying vulnerabilities.
4. Corroborate	Test pivotal claims against external evidence.	Verify the claims that could change diagnosis, management, safety, or assessment using authoritative and current sources.	Provide or approve evidence hierarchy; teach local applicability and currency checks.	Claim-source-verdict verification table.
5. Calibrate	Make reliance and confidence revision explicit.	Classify advice as accept, modify, reject, or escalate; compare with the initial judgment and explain any confidence change.	Probe whether change is evidence-driven; discuss over- and underreliance.	Reliance decision with calibrated rationale.

Phase	Instructional purpose	Student action	Educator role	Observable output
6. Consolidate	Restore accountable reasoning and test transfer.	Produce the final response in the learner's own reasoning chain, disclose AI use, reflect on errors, and solve a related case unaided.	Give process-focused feedback and schedule delayed transfer.	Final answer, reflection, and unaided transfer performance.

3.1.1. Phase 1: Commit

Students encounter a case with a decision that matches their level of training. Before accessing GenAI, each learner submits four items: a concise problem representation; a provisional diagnosis, priority, or action; the evidence supporting it; and a confidence rating. For pharmacy, the decision might be a medication-related problem and proposed intervention; for nursing, a priority assessment and escalation plan; for rehabilitation, a functional hypothesis and treatment priority; for dentistry, a diagnostic or management decision. The purpose is not to reward premature certainty. It is to preserve independent generation and make later influence observable.

3.1.2. Phase 2: Consult

Students consult a permitted GenAI system using a standard minimum record: tool and displayed model, access date, full prompt, relevant settings or retrieval mode, and full output. A shared prompt can support experimental comparability, whereas learner-generated prompts can support authentic AI literacy. In either format, the AI role is bounded. It may offer hypotheses, critique, explanations, or simulated feedback; it does not issue an unreviewed patient-care instruction. No identifiable or confidential patient information is entered into public systems.

3.1.3. Phase 3: Challenge

Students interrogate the response before looking for confirmatory sources. Useful questions include: What important alternative is missing? Which statement is most uncertain? What evidence would contradict this recommendation? Which assumption may not apply in our setting? Could the response reflect demographic, geographic, or language bias? What additional information would change the answer? The goal is to turn a single polished output into a contestable set of claims. Faculty should resist giving a universal 'perfect prompt'; the educational object is disciplined questioning, not prompt mimicry.

3.1.4. Phase 4: Corroborate

Not every sentence deserves equal checking. Students identify pivotal claims—those that could alter diagnosis, treatment, escalation, patient safety, or a graded conclusion—and verify them using an explicit evidence hierarchy. Suitable sources include current professional guidelines, national formularies, institutional protocols, systematic reviews, and primary research where appropriate. Search-result snippets, anonymous websites, and references supplied by the model are leads, not proof. A compact verification record should link each claim to a source, note currency and contextual fit, and assign a verdict: supported, contradicted, partly supported, or unresolved.

3.1.5. Phase 5: Calibrate

Students decide whether to accept, modify, reject, or escalate each consequential recommendation. They compare the final decision with their pre-AI commitment and rate confidence again. The explanation must identify the evidence responsible for change. A student who changes from correct to incorrect after faulty AI advice has overrelied; a student who refuses well-supported correct advice without a defensible reason has underrelied. A student who appropriately modifies a generally correct answer to fit a patient's contraindication or local policy demonstrates calibrated reliance. Early evidence suggests that structured educational exposure can improve trust calibration, although stronger evaluation is needed [29].

3.1.6. Phase 6: Consolidate

The student submits an accountable final response that integrates—not copies—the verified information and clearly discloses the role of AI. A short reflection identifies the most influential AI claim, an error or limitation discovered, and a rule for future cases. Immediate consolidation should be followed by a delayed, related case completed without AI. This transfer task is essential: it distinguishes genuine learning from temporary assisted performance.

3.2. Designing VS-CBL Cases

The educational value of VS-CBL depends on deliberate variation rather than random exposure to whatever a commercial model happens to generate. Faculty should create or adapt a bank of cases and pretest the expected outputs near the time of teaching. Cases should differ along at least five dimensions: AI accuracy, error subtlety, clinical or professional stakes, completeness of information, and learner expertise. Table 2 provides a design matrix.

Table 2. Case-design matrix for calibrated reliance.

Dimension	Illustrative levels	Educational function
AI accuracy	Accurate; partly accurate; incorrect; internally inconsistent	Prevents students from learning that AI is always wrong or always right; permits measurement of both over- and underreliance.
Error subtlety	Obvious factual error; plausible omission; outdated guidance; wrong population; fabricated reference; locally inapplicable advice	Moves verification beyond easy fact-checking toward epistemic and contextual judgment.
Stakes	Low (wording or recall); moderate (formative case decision); high (dose, contraindication, emergency escalation)	Allows required evidence and supervision to scale with potential harm.
Information completeness	Complete; deliberately incomplete; ambiguous; conflicting	Tests whether learners request missing data and tolerate unresolved uncertainty.
Learner expertise	Novice; intermediate; advanced	Supports fading of prompts, expansion of source choice, and increasing complexity.
Contextual variation	Local guideline; resource constraint; multilingual communication; cultural preference; scope-of-practice boundary	Tests global and local applicability rather than decontextualised correctness.

A sequence should include correct and incorrect AI advice in unpredictable proportions. If every teaching case contains a planted error, students may learn reflexive distrust rather than calibration. Conversely, if errors are rare and never consequential, learners may treat verification as ceremonial. The most instructive cases often contain a mixture: a broadly reasonable answer with one consequential omission, a real but outdated reference, or advice that is correct in one jurisdiction but not another.

3.3. A Feasible 90-minute Learning Session

A single facilitated session can introduce the cycle: orientation and privacy rules (10 minutes); unaided Commit (15 minutes); Consult and Challenge (20 minutes); Corroborate using an approved evidence set (15 minutes); Calibrate and final answer (10 minutes); small-group comparison and faculty debrief (15 minutes); and Consolidate reflection (5 minutes). A related unaided case can be assigned 7-14 days later. For large cohorts, students can work individually during Commit and in pairs during corroboration, while faculty use a common rubric and model answer. Structured GenAI facilitation has shown potential to support self-directed learning, while underscoring the importance of instructional design [30]. The framework can also be distributed across asynchronous learning, simulation, ward-based teaching, or interprofessional case conferences.

3.4. Assessment Blueprint

Assessment should separate the quality of the final answer from the quality of reliance. A learner may reach the right answer through unsafe acceptance or the wrong answer after a defensible response to ambiguous evidence. Both product and process matter. Table 3 proposes outcomes that can be used formatively and, after validation, in research or programme evaluation.

Table 3. Assessment blueprint for VS-CBL.

Outcome	Operational indicator	Interpretive purpose
Unaided task performance	Accuracy or rubric score recorded before AI exposure	Baseline reasoning and later independent transfer
AI-assisted task performance	Final accuracy or rubric score after the cycle	Whether the complete learning design improves the immediate product
Overreliance	Proportion of initially correct or undecided responses changed to an incorrect response after incorrect AI advice; or acceptance of planted incorrect advice	Susceptibility to unsafe automation influence

Outcome	Operational indicator	Interpretive purpose
Underreliance	Proportion of useful correct advice rejected without a defensible evidence-based reason	Failure to benefit from appropriate assistance
Appropriate reliance	Correct acceptance plus correct rejection, interpreted with task context and justification	Alignment between reliance and advice quality
Confidence calibration	Association or gap between confidence and accuracy; Brier score for binary judgments where appropriate	Metacognitive monitoring rather than confidence alone
Verification quality	Rubric for claim prioritisation, source authority, currency, triangulation, and local applicability	Whether checking is substantive and proportionate
Reasoning and justification	Key-features or analytic rubric applied to problem representation, evidence use, alternatives, and escalation	Accountable clinical or professional reasoning
Delayed unaided transfer	Performance on a novel but structurally related case without GenAI	Learning retained beyond assisted performance
Implementation outcomes	Time, workload, fidelity, acceptability, access barriers, and privacy incidents	Feasibility and equity of scale-up

For experimental evaluation, an appropriate comparator should reflect the causal question. VS-CBL could be compared with unguided GenAI-assisted case learning and conventional case-based learning. Assessors should be blinded where feasible, and cases should be counterbalanced because a particular AI error can vary in detectability. Primary outcomes should be prespecified. Delayed unaided transfer and overreliance are stronger candidates than satisfaction. Model name, version or displayed label, retrieval capability, temperature if controllable, system prompts if known, access date, and prompts should be reported because outputs can change over time.

3.5. Faculty Facilitation, Governance, and Safety

VS-CBL does not make faculty optional. It changes the faculty role from sole content transmitter to designer of epistemic conditions, verifier of case safety, and coach of reasoning. Before a session, faculty should confirm the case's learning objectives, generate and archive expected AI responses, identify pivotal claims and authoritative sources, rate possible harms, and prepare debrief questions. During debriefing, they should ask how the learner knew, what changed the decision, and what remained uncertain—not merely reveal whether the AI was correct.

Governance should define permitted systems, data handling, disclosure, assessment rules, accessibility, and escalation. Public GenAI systems must not receive identifiable patient or learner data unless an institution has explicitly approved that use. High-stakes clinical decisions require qualified human supervision and current local protocols. If a model's behaviour changes materially, cases should be revalidated. Students should be told when output has been deliberately altered or simulated for teaching after the exercise, so that deception does not undermine trust.

Ethical AI education also requires avoiding a hidden curriculum in which students are warned against tools that faculty use privately. Transparent modelling is preferable: educators can show a documented interaction, identify what they verified, disclose uncertainty, and demonstrate rejection of a persuasive but unsupported claim. Ethical and assessment guidance in health professions education emphasises human accountability, disclosure, privacy, validity, and governance rather than technology-led adoption [31, 32].

3.6. International and Resource-diverse Implementation

A globally acceptable framework cannot assume continuous broadband, paid model access, English-only content, or universal guidelines. VS-CBL is product-agnostic and can be implemented with institutionally archived outputs when live access is unreliable. In an offline version, all learners receive the same pre-generated AI response and complete Challenge, Corroborate, Calibrate, and Consolidate phases using printed or locally stored sources. This preserves most of the mechanism while improving experimental control and equity.

Verification must privilege locally authoritative evidence. International guidance can be useful, but it may assume medicines, diagnostics, referral pathways, professional roles, or patient preferences unavailable in a particular setting. Cases should therefore include local formularies and protocols alongside international sources. Multilingual outputs require verification of clinical meaning, tone, and cultural fit—not only grammatical quality. Students should be invited to identify whose population, resources, and values the AI response assumes.

Resource limitation is not only an implementation barrier; it is an educational variable. A recommendation that is technically optimal in a tertiary centre may be unsafe or impossible in a rural clinic. Calibrated reliance includes recognising that contextual mismatch. Programmes should also monitor whether differences in device ownership, digital literacy, disability access, or subscription level create unequal opportunity. AI literacy should be taught as a professional capability with supported access, not treated as an extracurricular advantage [33-36].

4. Discussion

4.1. What VS-CBL Adds to Existing Approaches

Existing GenAI curricula commonly emphasise prompt writing, awareness of hallucination, ethical principles, source checking, or comparison of AI and expert answers. Recent scholarship includes a pharmacology-specific source-checking process and a cross-profession scoping review with implementation and evaluation tools [37, 38]. VS-CBL links verification to a time-sensitive reliance architecture. Its distinctive features are: an observable pre-AI commitment; adversarial challenge before evidence search; prioritised corroboration of consequential claims; an explicit accept-modify-reject-escalate decision; paired confidence measurement; and delayed unaided transfer. These features make the mechanism and outcomes of reliance measurable.

The framework also rejects a false choice between teaching with AI and teaching about AI. Students learn disciplinary content through cases while simultaneously learning how AI changes evidence search, confidence, and judgment. This integration matters because decontextualised AI-literacy lectures may not transfer to a pressured clinical or professional decision. Conversely, using GenAI for content learning without discussing reliance can normalise invisible dependence.

4.2. Limitations

VS-CBL is a proposed framework, not an evaluated intervention. Its six phases are theoretically and empirically informed, but their sequence, dosage, and independent contribution have not been established. The critical synthesis was focused and purposive rather than systematic, so relevant literature may have been missed and no certainty-of-evidence judgments were made. Evidence about cognitive forcing and unguarded GenAI includes studies outside health professions education; transfer of those findings should be tested rather than assumed. The framework also requires faculty time for case validation and debriefing. Commercial model instability may reduce reproducibility, while tightly scripted outputs may reduce ecological validity.

Measurement remains challenging. Appropriate reliance cannot always be reduced to agreement with a gold standard because clinical evidence can be uncertain and contextual values legitimate. Confidence scales are sensitive to individual and cultural response patterns. Process logs may change behaviour, and a planted error may be easier to detect than an organically generated one. Finally, no framework can eliminate the need for institutional governance or qualified supervision in patient-facing use.

4.3. Research Agenda

The immediate priority is prospective validation. A three-arm randomized trial could compare VS-CBL, unguided GenAI-assisted case-based learning, and conventional case-based learning. Prespecified primary outcomes should include delayed unaided clinical-reasoning performance and overreliance on incorrect AI advice. Secondary outcomes could include immediate aided performance, verification quality, confidence calibration, cognitive load, time, and acceptability. The design should separate exposure to AI from the verification scaffold and should include correct as well as incorrect advice to measure both over- and underreliance.

Multicentre studies across professions, languages, and income settings are necessary before claims of generalisability. Factorial or micro-randomized designs could test which phases are essential—for example, precommitment, adversarial challenge, or required source verification. Longitudinal work should examine whether the scaffold can fade without losing safe behaviour, whether skills transfer across models and clinical domains, and whether calibrated reliance persists in simulation or workplace performance. Qualitative studies should investigate how learners experience uncertainty, authority, effort, and professional responsibility. Economic and implementation studies should quantify faculty workload and identify feasible low-connectivity adaptations.

Existing reviews report promising effects but substantial heterogeneity, limited learning-theory use, risk of bias, and little evidence at behavioural or patient levels [1-4, 28]. Future trials should therefore report intervention components precisely, use validated or transparently developed measures, register protocols, publish prompts and de-identified outputs where permitted, and distinguish AI-assisted performance from unaided learning. Because model behaviour evolves, replication should be conceptual—testing the scaffold's mechanisms across tools—rather than tied permanently to one product.

5. Conclusion

Health professions students do not need a rule that GenAI is trustworthy or untrustworthy. They need disciplined practice deciding whether a particular claim deserves reliance in a particular context. VS-CBL makes that practice teachable by protecting an unaided judgment, requiring active challenge, prioritising corroboration, and making acceptance, modifica-

tion, rejection, or escalation explicit. It also restores the criterion that matters for education: what the learner can reason and do when assistance is absent.

The framework is intentionally product-agnostic and adaptable across professions and resource settings. Its novelty lies in translating appropriate-reliance theory into a complete case-based instructional and assessment sequence. It should now be treated as a testable proposal. Rigorous prospective evaluation must determine whether the added verification work improves calibration and durable learning enough to justify its cost.

6. Recommendations

Health professions programmes should introduce VS-CBL through supervised, low-stakes case-based activities; require an unaided commitment before AI consultation; assess challenge and verification quality, confidence calibration, and delayed unaided transfer; prohibit identifiable data in unapproved systems; and validate cases and model outputs against current local guidance. Institutions should provide equitable access, faculty development, and governance for permitted tools and disclosure. Researchers should evaluate the framework prospectively across professions and resource settings, reporting model versions, prompts, implementation fidelity, workload, and both overreliance and underreliance.

Abbreviations

AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
VS-CBL	Verification-Scaffolded Case-Based Learning

Author Contributions

Adesegun Nurudeen Osijirin: Conceptualization, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing

Shamsuddeen Mohammed Sada: Conceptualization, Validation, Writing – review & editing

Utibe Edmond Victor: Resources, Validation, Writing – original draft, Writing – review & editing

Oka Obasi: Methodology, Writing – review & editing

Nsangaette Asuquo Udo: Methodology, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Pham TD, Karunaratne N, Exintaris B, Liu D, Lay T, Yuriev E, et al. The impact of generative AI on health professional education: a systematic review in the context of student learning. Med Educ. 2025; 59(12): 1280-1289. https://doi.org/10.1111/medu.15746

[2] Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024; 46(4): 446-470. https://doi.org/10.1080/0142159X.2024.2314198

[3] Lai NM, Lim YS, Win MT, Bhargava P, Thomas P, Ong QC. The effectiveness of artificial intelligence in undergraduate health professions education: a systematic review and meta-analysis of randomized controlled trials. JMIR Med Educ. 2026; 12: e88933. https://doi.org/10.2196/88933

[4] Abidi SH, Almazan J, Fabiyi O, Zehra F, Tariq M. AI-supported case-based learning in medical education: a comprehensive scoping review. Front Med (Lausanne). 2026; 13: 1798097. https://doi.org/10.3389/fmed.2026.1798097

[5] Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023; 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274

[6] Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024; 7(10): e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

[7] Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025; 101(1195): 458-463. https://doi.org/10.1093/postmj/qgae170

[8] Bastani H, Bastani O, Sungu A, Ge H, Kabakcı Ö, Mariman R. Generative AI without guardrails can harm learning: evidence from high school mathematics. Proc Natl Acad Sci U S A. 2025; 122(26): e2422633122. https://doi.org/10.1073/pnas.2422633122

[9] Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med. 2021; 4: 31. https://doi.org/10.1038/s41746-021-00385-9

[10] Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017; 24(2): 423-431. https://doi.org/10.1093/jamia/ocw105

[11] Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004; 46(1): 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

[12] Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors. 2015; 57(3): 407-434. https://doi.org/10.1177/0018720814547570

[13] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023; 55(12): 248. https://doi.org/10.1145/3571730

[14] Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. 2023; 15(2): e35179. https://doi.org/10.7759/cureus.35179

[15] Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform. 2024; 12: e54345. https://doi.org/10.2196/54345

[16] Risko EF, Gilbert SJ. Cognitive offloading. Trends Cogn Sci. 2016; 20(9): 676-688. https://doi.org/10.1016/j.tics.2016.07.002

[17] Sparrow B, Liu J, Wegner DM. Google effects on memory: cognitive consequences of having information at our fingertips. Science. 2011; 333(6043): 776-778. https://doi.org/10.1126/science.1207745

[18] Fisher M, Goddu MK, Keil FC. Searching for explanations: how the Internet inflates estimates of internal knowledge. J Exp Psychol Gen. 2015; 144(3): 674-687. https://doi.org/10.1037/xge0000070

[19] Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc ACM Hum-Comput Interact. 2021; 5(CSCW1): Article 188. https://doi.org/10.1145/3449287

[20] Croskerry P. Cognitive forcing strategies in clinical decisionmaking. Ann Emerg Med. 2003; 41(1): 110-120. https://doi.org/10.1067/mem.2003.22

[21] Mamede S, van Gog T, van den Berge K, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. JAMA. 2010; 304(11): 1198-1203. https://doi.org/10.1001/jama.2010.1276

[22] Zimmerman BJ. Becoming a self-regulated learner: an overview. Theory Pract. 2002; 41(2): 64-70. https://doi.org/10.1207/s15430421tip4102_2

[23] Sandars J, Cleary TJ. Self-regulation theory: applications to medical education. AMEE Guide No. 58. Med Teach. 2011; 33(11): 875-886. https://doi.org/10.3109/0142159X.2011.595434

[24] Garbayo LS, Harris DM, Fiore SM, Robinson M, Kibble JD. A metacognitive confidence calibration tool to help medical students scaffold diagnostic reasoning in decision-making during high-fidelity patient simulations. Adv Physiol Educ. 2023; 47(1): 71-81. https://doi.org/10.1152/advan.00156.2021

[25] Thistlethwaite JE, Davies D, Ekeocha S, et al. The effectiveness of case-based learning in health professional education: a BEME systematic review. BEME Guide No. 23. Med Teach. 2012; 34(6): e421-e444. https://doi.org/10.3109/0142159X.2012.680939

[26] McLean SF. Case-based learning and its application in medical and health-care fields: a review of worldwide literature. J Med Educ Curric Dev. 2016; 3: JMECD. S20377. https://doi.org/10.4137/JMECD.S20377

[27] Varma B, Karuveettil V, Fernandez R, et al. Effectiveness of case-based learning in comparison to alternate learning methods on learning competencies and student satisfaction among healthcare professional students: a systematic review. J Educ Health Promot. 2025; 14: 76. https://doi.org/10.4103/jehp.jehp_510_24

[28] Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ. 2025; 25: 129. https://doi.org/10.1186/s12909-025-06719-5

[29] Sakamoto T, Harada Y, Shimizu T. Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for medical students: quasi-experimental study. JMIR Form Res. 2024; 8: e58666. https://doi.org/10.2196/58666

[30] Wang S, Zuo Y, Zou B, Liu G, Zhou J, Zheng Y, et al. Enhancing self-directed learning with custom GPT facilitation: a randomized controlled trial. Med Teach. 2025; 47(7): 1126-1133. https://doi.org/10.1080/0142159X.2024.2413023

[31] Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach. 2023; 45(6): 574-584. https://doi.org/10.1080/0142159X.2023.2186203

[32] Masters K, MacNeil H, Benjamin J, et al. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. Med Teach. 2025. https://doi.org/10.1080/0142159X.2024.2445037

[33] Preiksaitis C, Rose C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. JMIR Med Educ. 2023; 9: e48785. https://doi.org/10.2196/48785

[34] Basil M, Ahmed W, Hajeomar R, Strawbridge J, Lynch M, Mukhalalati B. A scoping review of the use of generative artificial intelligence tools in health profession education. BMC Med Educ. 2026; 26: 291. https://doi.org/10.1186/s12909-025-08527-3

[35] Boman M, Jhun P, Schaekermann M, et al. Scaffolding for success: blending learning with and about generative AI in medical education. Med Teach. Published online 2025. https://doi.org/10.1080/0142159X.2025.2571041

[36] Naamati-Schneider L, Alt D. Developing AI literacy in medical and health professions education: exploring ChatGPT, ChatBots development, and simulation-based learning. Nurse Educ Pract. 2025; 88: 104572. https://doi.org/10.1016/j.nepr.2025.104572

[37] Che X, Liu Y. Guiding student use of generative AI in undergraduate pharmacology: a narrative review and proposed source-checking process framework. Front Med (Lausanne). 2026; 13: 1910543. https://doi.org/10.3389/fmed.2026.1910543

[38] Astbury H, Fortune E, Ponte CD, Lyons K, Shaw L. Generative artificial intelligence for teaching and assessment in health professions education: a scoping review. Nurse Educ Pract. 2026; 91: 104697. https://doi.org/10.1016/j.nepr.2025.104697