

Clinical Reasoning in the AI era: A focused scoping review and narrative synthesis of conceptual, pedagogical, and assessment frameworks in health professions education

Naomi Campbell-Woods

`naomi.campbell-woods@health.qld.gov.au`

Queensland Children's Hospital

Research Article

Keywords: Clinical reasoning, artificial intelligence, generative AI, health professions education, assessment, medical education, scoping review

Posted Date: July 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9918525/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Clinical reasoning is a core aim of health professions education, yet there is still no stable agreement on what the construct includes, how it should be taught, or how it should be assessed. This problem has become more urgent with the rapid uptake of artificial intelligence in medical and allied health education. AI studies increasingly claim to improve clinical reasoning, but they often operationalise it through narrow proxies such as diagnostic accuracy, history-taking completeness, key feature performance, or documentation quality rather than through an explicit construct model (Daniel et al., 2019; Durning et al., 2013; Gordon et al., 2022; Masters et al., 2025). This scoping review mapped how clinical reasoning is conceptualized, how pedagogical and assessment frameworks operationalise it, and how AI-enabled educational studies inherit, adapt, or narrow those frameworks.

Methods

A focused scoping review was conducted using two concept-based deep searches of the scholarly literature completed on May 7 and May 10, 2026. One search targeted empirical studies on generative AI and clinical reasoning in health professions education. The second targeted conceptual, curricular, and assessment frameworks relevant to clinical reasoning in the AI era. Together, the searches yielded 158 ranked records. Records were reviewed for direct relevance to clinical reasoning conceptualization, assessment, pedagogy, or AI-enabled implementation in health professions education. The synthesis centered on a purposive analytic corpus of framework papers, assessment-method papers, systematic reviews, and empirical AI studies, with medicine as the dominant case and allied health included when concepts or measures transferred across professions.

Results

The literature separated into three linked bodies of work. First, conceptual papers described clinical reasoning through script theory, dual-process and metacognitive models, and situated cognition, while more recent assessment papers emphasized longitudinal and programmatic approaches rather than single-test solutions (Charlin, Tardif, et al., 2000; Guth et al., 2024; Lubarsky et al., 2015; Pelaccia et al., 2011; Rencic et al., 2020b; Torre et al., 2024). Second, the assessment literature offered a broad toolbox, including key feature problems, script concordance tests, history-taking indicators, oral presentation tools, post-encounter forms, documentation rubrics, simulation, and workplace-based assessment, but repeatedly warned that no single instrument can capture the construct in full (Daniel et al., 2019; Schuwirth et al., 2020; Thampy et al., 2019). Third, AI-enabled education studies mostly attached these established tools to simulations, automated feedback, note scoring, and diagnostic support systems rather than introducing a new theory of clinical reasoning (Brügge et al., 2024; Çiçek et al., 2024; Dai et al., 2026; Hudon et al., 2025; Roemer et al., 2026; Saloojee et al., 2026; Schaye et al., 2024; Sur et al.,

2026). Reported gains were most consistent in case-specific and highly scoreable domains such as data gathering, illness-script quality, diagnostic justification, and documentation structure, whereas richer dimensions of reasoning including uncertainty management, contextual adaptation, metacognition, and collaborative human-AI judgment remained thinly assessed (Abdulnour et al., 2025; Plackett et al., 2022; Wu, 2025).

Conclusions

The literature suggests the major challenge may no longer be the availability of measures, but the absence of an AI-era integrate construct model that links what should be taught, what should be protected from off-loading, and what should count as evidence of sound reasoning. Current AI studies mostly measure the fragments of reasoning that are easiest to observe, script, score, or automate. The next step for the field is not another isolated AI intervention trial, but a theory-informed framework that integrates knowledge organization, process, uncertainty, context, sociotechnical judgment, and programmatic assessment design.

Background

Clinical reasoning sits near the center of safe practice across medicine and the health professions, yet it remains one of the least settled constructs in the education literature. Gordon and colleagues described a fractured assessment literature marked by inconsistent terminology, weak conceptual framing, and uneven validity reporting (Gordon et al., 2022). Durning and colleagues similarly argued that clinical reasoning can be framed as cognition, behavior, performance, social interaction, or situated adaptation, depending on the theoretical lens being used (Durning et al., 2013). The result is a mismatch between educational ambition and measurement practice.

That mismatch matters more now because AI has entered the learning environment faster than the construct has been stabilized. AI tools can simulate patients, generate feedback, score notes, build script concordance tests, draft key feature problems, and function as live diagnostic aids (Abouzeid & Sallam, 2025; Holderried et al., 2024; Hudon et al., 2025; Roemer et al., 2026; Zafar et al., 2025). These developments create a temptation to treat whatever AI can score reliably as the educational target itself.

The pre-AI literature already warned against one-dimensional assessment. Daniel and colleagues mapped a broad menu of clinical reasoning assessments and argued for programs of assessment built from complementary methods rather than reliance on a single tool (Daniel et al., 2019). Schuwirth and colleagues described a shift from single-problem testing toward more situated and judgment-based views of performance (Schuwirth et al., 2020). Rencic and colleagues argued that valid assessment must take context seriously because clinical reasoning performance emerges from interactions among learner, patient, task, rater, method, and environment (Rencic et al., 2020b, 2020a).

The question for this review was therefore not simply whether AI helps. The more important question was how the literature conceptualises clinical reasoning, how pedagogical frameworks operationalise it, and whether AI studies are expanding the construct or merely narrowing it to what is easiest to automate. The review had three aims:

- to map the main conceptual frameworks used to describe clinical reasoning in health professions education
- to identify the assessment and pedagogical frameworks used to operationalise those constructs
- to examine how AI-enabled educational studies position themselves in relation to that prior literature

Methods

Design and scope

This study was conducted as a focused scoping review with a narrative synthesis of findings. The review mapped conceptual, pedagogical, and assessment literature on clinical reasoning, then related that map to empirical and methodological AI literature in health professions education. Medicine was the dominant index principle, while nursing, occupational therapy, physiotherapy, dentistry, and related professions were incorporated where the constructs or assessment approaches were conceptually relevant (Gordon et al., 2022; Huang, 2026; Rodrigues-Pereira et al., 2025; Sigirtmac et al., 2026).

Information sources and search strategy

Two concept-based deep searches of the scholarly literature were used across MEDLINE, Embase, CINAHL, ERIC, PsycINFO, Scopus and Web of Science. Grey literature searching included professional organization reports, citation chaining and hand searching of key medical education journals. Given the rapidly evolving nature of generative AI literature, relevant pre-print repositories including medRxiv and arXiv were also reviewed.

The first search completed targeted empirical studies of generative AI and clinical reasoning in health professions education and yielded 33 ranked records. The subsequent search targeted papers that conceptualised clinical reasoning, described pedagogical or assessment frameworks, or linked those frameworks to AI in medical education and yielded 125 ranked records.

Eligibility criteria and synthesis

Papers were eligible if they conceptualised clinical reasoning explicitly, described or validated an assessment or pedagogical framework, reported an AI-enabled intervention or supervision framework directly linked to reasoning, or provided review-level evidence relevant to clinical reasoning pedagogy or

assessment. Records were reviewed using metadata, abstracts, and full-text PDF’s. Data were charted for profession, conceptual framework, assessment family, target construct, AI role, and stated limitations. The synthesis was organized in three layers: conceptual frameworks, assessment and pedagogical frameworks, and AI-enabled studies that inherited, adapted, or narrowed those prior frameworks.

Results

Search overview

The two completed searches together yielded 158 ranked records. The empirical search concentrated on the current GenAI intervention literature. The broader frameworks search added the conceptual and assessment literature needed to interpret those intervention studies (Table 1).

Table 1
Search Results

Search	Focus	Records returned	Main contribution
GenAI and clinical reasoning in health professions education	Empirical AI studies	33	Intervention designs, learner groups, and measured outcomes
Clinical reasoning frameworks for AI in medical education	Conceptual and assessment literature plus AI-linked frameworks	125	Theoretical anchors, assessment families, and AI-era supervision and validity arguments

Clinical reasoning is supported by mature but non-identical frameworks

The broader literature did not show a lack of theory. It showed several mature theories emphasizing different parts of the same educational target (Table 2).

Script theory remains the clearest bridge between knowledge organisation and diagnostic performance. Illness scripts are structured networks linking enabling conditions, faults, and consequences, and education in this tradition emphasizes case comparison, self-explanation, and script refinement (Charlin, Tardif, et al., 2000; Lubarsky et al., 2015). This is also the clearest bridge into AI studies that scaffold illness-script development or script concordance tasks (Hudon et al., 2025; Sur et al., 2026).

Dual-process accounts frame reasoning as interaction between rapid pattern recognition and slower analytic checking (Croskerry, 2009; Pelaccia et al., 2011). The educational implication is that assessment cannot focus only on final answers. It also needs ways to detect prioritization, bias, and reflective checking (Thammasitboon et al., 2018; Thampy et al., 2019).

Situated cognition pushes the construct outward. Clinical reasoning cannot be understood purely as a mental process inside the learner because it emerges through interactions among patient, physician, environment, and task (Durning et al., 2013; Rencic et al., 2020b, 2020a). Programmatic assessment extends this argument by emphasizing broad sampling, triangulation, proportionality between evidence and stakes, and longitudinal interpretation rather than single scores (Guth et al., 2024; Torre et al., 2024).

Table 2
Clinical reasoning frameworks

Framework	Core idea	Educational implication
Script theory	Reasoning depends on organized illness scripts	Teach and assess knowledge organization and script refinement
Dual-process and metacognitive models	Reasoning emerges from interaction between pattern recognition and analytic checking	Assess prioritization, bias monitoring, and reflective checking, not only final answers
Situated cognition	Reasoning performance is context dependent and emerges from learner, patient, task, rater, and environment	Sample broadly across contexts and avoid single-method claims
Programmatic assessment	No single instrument can support high-stakes claims about reasoning	Build longitudinal systems that triangulate multiple data sources and match evidence to stakes
AI-era sociotechnical views	Reasoning now includes supervision, appraisal, and appropriate use of AI	Assess human-AI collaboration rather than unaided cognition alone

The assessment literature offers many tools, but each captures only a fragment of the construct

The assessment literature was rich and highly practical, but explicit about its limits. Key feature problems focus on critical decisions (Hrynchak et al., 2014; Page et al., 1995). Script concordance tests assess reasoning under uncertainty (Charlin, Roy, et al., 2000; Lubarsky et al., 2013). Observable process tools such as CRI-HTI, ART, and REACT focus on enacted behaviors including hypothesis-directed questioning, problem representation, differential diagnosis, management, and reflection (Brügge et al., 2024; Peterson et al., 2022; Thammasitboon et al., 2018). Documentation-based tools such as post-encounter forms and the Revised-IDEA family assess diagnostic justification in written notes (Schaye et al., 2021; Schaye et al., 2022; Schaye et al., 2024). Virtual patients and simulation support repeated practice and structured feedback, but remain surrogates for the clinical world (Hege et al., 2018; Plackett et al., 2022; Waechter et al., 2026). Across these families, the same conclusion recurred: no single method captures clinical reasoning in full (Daniel et al., 2019; Schuwirth et al., 2020; Thampy et al., 2019).

AI studies mostly automate teaching, feedback, and scoring of pre-existing reasoning constructs

The empirical AI literature did not introduce a new dominant theory of clinical reasoning. Instead, it mainly attached AI to existing targets (Table 3).

One cluster uses AI as a simulated patient plus feedback engine for history taking. ChatGPT-supported patient simulation plus immediate structured feedback improved CRI-HTI scores (Brügge et al., 2024), and GPT-4 produced history-taking feedback with high agreement with a human rater (Holderried et al., 2024).

A second cluster uses AI to generate or deliver feedback around established written formats. ChatGPT-generated feedback was broadly comparable to expert feedback on key-feature learning overall, but weaker for more complex cases (Çiçek et al., 2024). Similar crossover work found no meaningful advantage of personalized AI feedback over expert comments (Nissen et al., 2025).

A third cluster uses AI to scaffold illness-script development or uncertainty-based reasoning directly. Microsoft Copilot improved illness-script quality on the ISAT (Sur et al., 2026), and AI-authored and AI-scored script concordance tests were feasible and educationally useful (Hudon et al., 2025).

A fourth cluster uses AI to assess or coach reasoning through written documentation. Machine learning and large language model approaches have shown promise for scoring diagnostic justification and reasoning documentation in essays and notes (Cianciolo et al., 2021; Schaye et al., 2022; Schaye et al., 2024). These studies are important because they make reasoning traces more scalable, but they assess written reasoning documentation rather than the whole reasoning process.

A fifth cluster uses AI as a live diagnostic aid or performance support tool. LLM support improved vignette-based diagnostic accuracy and confidence in some studies (Roemer et al., 2026), but not all (Goh et al., 2024), and self-directed bedside use did not improve overall real-patient exam performance (Saloojee et al., 2026).

A sixth cluster uses AI-enhanced simulation in broader performance environments. ChatGPT feedback improved decision accuracy and related outcomes in VR-based OSCE work (Dai et al., 2026), while fully automated AI-generated questions and feedback improved structured video-exam performance in early learners (Kiyak et al., 2025). Allied health studies showed the same pattern: effects were more plausible when AI use was tightly scaffolded to a named framework and weaker when use was looser (Huang, 2026; Sigirtmac et al., 2026).

Table 3
Construct and limitations of measuring clinical reasoning

AI use case	Typical construct	Main limitation
Simulated patient plus feedback	History-taking reasoning process	Narrow focus on observable dialogue
AI feedback on written case questions	Critical decision making	Complex cases still need expert nuance
Illness-script and uncertainty support	Script quality and uncertainty reasoning	Narrow construct slice and limited transfer evidence
Documentation scoring	Written diagnostic justification	Documentation is not equivalent to full reasoning
Diagnostic aid during case solving	Diagnostic output and calibration	Often outcome focused and vignette bound
Immersive or clinical performance support	Bundled encounter performance	Reasoning becomes diffuse inside larger performance scores

The strongest gains appear in the parts of reasoning that are easiest to structure and score

Across both the virtual patient and AI literatures, the most consistent gains occurred in case-specific, bounded, and scoreable domains. Virtual patient reviews found more reliable improvement in data gathering, ideas about diagnosis, and patient management than in general problem-solving (Plackett et al., 2022). The same pattern appears in AI studies. History-taking systems improve structured information gathering (Brügge et al., 2024; Holderried et al., 2024). Script-based systems improve script quality (Hudon et al., 2025; Sur et al., 2026). Documentation models score explicit differential diagnoses and diagnostic explanations more reliably than they assess deeper interpretive reasoning (Schaye et al., 2022; Schaye et al., 2024). AI appears most useful when the target construct is already decomposed into visible steps, bounded written outputs, or clearly specified feedback categories.

Under-measured dimensions in the AI literature

Several dimensions described in the conceptual literature remain thinly assessed in AI studies (Table 4).

Table 4
Gaps in the Measurements Used in Studies

Under-measured dimension	Main gap in AI studies
Knowledge organization	Few studies connect gains to longitudinal script formation outside short tasks
Process tracing	Internal hypothesis revision and cue weighting are rarely captured directly
Uncertainty handling	Limited use in authentic or longitudinal settings
Contextual adaptation	Most AI studies remain decontextualized, brief, and single site
Sociotechnical judgment	Few validated tools for collaborative human-AI reasoning

Discussion

The central finding of this review is not that clinical reasoning lacks theory, nor that AI studies lack outcomes. The problem is the missing bridge between them. Health professions education already has mature traditions for understanding clinical reasoning as knowledge organisation, dual-process cognition, contextual performance, and longitudinal professional growth (Charlin, Tardif, et al., 2000; Pelaccia et al., 2011; Rencic et al., 2020b; Torre et al., 2024). At the same time, AI studies increasingly measure reasoning through local proxies that are convenient for simulation, feedback, or automation (Brügge et al., 2024; Çiçek et al., 2024; Roemer et al., 2026; Schaye et al., 2024). What the field still lacks is an explicit AI-era construct model linking pedagogy, assessment, supervision, and interpretation.

This matters because teaching cannot be reliable when the target is unstable. If clinical reasoning is treated as script formation, a curriculum should emphasize contrastive cases, self-explanation, and uncertainty-sensitive assessment (Charlin, Roy, et al., 2000; Lubarsky et al., 2015). If it is treated as situated performance, the same curriculum should rely on longitudinal sampling, expert judgment, and triangulation rather than single written tests (Guth et al., 2024; Rencic et al., 2020b; Torre et al., 2024). If it now includes sociotechnical judgment, then learners must also demonstrate when to defer, when to challenge AI, and how to preserve independent problem representation before consulting a model (Abdulnour et al., 2025; Wu, 2025). Without explicit alignment across these layers, AI trials will continue to generate local outcome gains that are hard to interpret.

Current optimism about AI-enhanced teaching is also partly a function of measurement choice. Gains are strongest where reasoning has already been reduced to visible traces, such as asking focused questions, selecting key decisions, writing a structured note, or building an illness script within a bounded template (Brügge et al., 2024; Schaye et al., 2022; Schaye et al., 2024; Sur et al., 2026). These are valuable achievements. They are not the same as demonstrating robust reasoning under uncertainty, distraction, interpersonal complexity, or shifting workplace context. The bedside trial by Saloojee and

colleagues is instructive because the move from simulation to real patient care largely erased the expected performance advantage of tool access (Saloojee et al., 2026).

Several implications follow.

- Clinical reasoning should be taught explicitly rather than assumed to emerge from apprenticeship or tool exposure alone (Cooper et al., 2020; Durning et al., 2024).
- Assessment should be designed programmatically. Automated scores may be useful for formative coaching, but high-stakes claims still require broader and better-quality evidence (Guth et al., 2024; Masters et al., 2025; Torre et al., 2024).
- AI supervision needs to preserve productive struggle. De-skilling, never-skilling, and mis-skilling are real educational risks when learners off-load core reasoning tasks too early (Abdulnour et al., 2025).
- The same construct problem appears across health professions, not only in medicine (Dai et al., 2026; Huang, 2026; Sigirtmac et al., 2026).

Proposed framework

The evidence map suggests that a useful AI-era framework should include six linked dimensions:

- knowledge organization
- process enactment
- uncertainty handling
- contextual performance
- sociotechnical judgment
- programmatic architecture

The first five dimensions describe the construct itself. The sixth specifies how evidence about the construct should be assembled and interpreted over time.

The figure positions clinical reasoning as a multidimensional construct shaped by pedagogical inputs and interpreted through a programmatic assessment system (Fig. 1). Five dimensions describe the reasoning construct itself: knowledge organization, process enactment, uncertainty handling, contextual performance, and sociotechnical judgment. A sixth dimension, programmatic architecture, specifies how evidence should be assembled and interpreted through triangulation, proportionality, longitudinal sampling, coaching, and narrative judgment. The model proposes that AI should not be treated as a separate educational topic or as a simple add-on tool. Instead, AI alters the conditions under which reasoning is learned, enacted, observed, and judged. These areas are further elaborated on below (Table 5).

Table 5
Suitable assessments for clinical reasoning dimensions and risks

Dimension	Core question	What should be taught	Suitable assessment examples	Main risk if omitted
Knowledge organization	How is knowledge structured for recognition and comparison of clinical problems	Illness script formation, semantic qualifiers, contrastive case comparison, self-explanation	ISAT, SCT, diagnostic justification tasks (Hudon et al., 2025; Lubarsky et al., 2015; Sur et al., 2026)	Learners may appear fluent while relying on shallow pattern matching or borrowed outputs
Process enactment	How does the learner move from cues to a justified working diagnosis or plan	Hypothesis-directed data gathering, problem representation, differential diagnosis, justification, revision	CRI-HTI, ART, REACT, post-encounter forms, structured reflection (Brügge et al., 2024; Goh et al., 2024; Peterson et al., 2022; Thammasitboon et al., 2018)	Correct answers may mask weak reasoning pathways
Uncertainty handling	How does the learner reason when information is incomplete, ambiguous, or conflicting	Calibration, reflective checking, bias recognition, tolerance of ambiguity	SCTs, uncertainty-focused key features, calibration measures (Charlin, Roy, et al., 2000; Hudon et al., 2025; Roemer et al., 2026)	AI may reinforce overconfidence, passive acceptance, or premature closure
Contextual performance	How does reasoning change across patients, tasks, environments, and levels of authenticity	Adaptation to case complexity, workflow, time pressure, team input	OSCEs, workplace-based assessment, repeated sampling across cases, bedside exams (Guth et al., 2024; Rencic et al., 2020b; Saloojee et al., 2026)	Decontextualized performance may be mistaken for real-world competence
Sociotechnical judgment	How does the learner use AI without off-loading core reasoning	Human-first problem representation, AI appraisal, verification, disclosure, supervised use	DEFT-AI style supervision, collaborative reasoning tasks, reliance measures (Abdulnour et al., 2025; Wu, 2025)	De-skilling, never-skilling, mis-skilling, and unsafe automation dependence
Programmatic architecture	How should evidence be assembled for educational decisions	Triangulation, proportionality, longitudinal interpretation, coaching	Programmatic assessment systems using mixed methods and staged decision making (Guth et al.,	Narrow automated scores may be over-interpreted in high-stakes decisions

Dimension	Core question	What should be taught	Suitable assessment examples	Main risk if omitted
			2024; Torre et al., 2024)	

Strengths and limitations

This review combined conceptual, assessment, and AI-enabled empirical literature rather than treating them as separate silos. It also included allied health studies where the frameworks were transferable.

This review also has important limitations. It was based on two concept-based deep searches and project-level full-text availability rather than a conventional manually exported multi-database search. Duplicate counts were not separately available from the project interface. The evidence base for AI-enabled studies remains small, recent, and often single-site. Several influential AI papers are preprints or early implementation studies. Some of the strongest recent arguments for programmatic assessment were available through a limited number of key papers rather than a dedicated search focused solely on programmatic assessment. As a scoping review, this study did not include formal risk-of-bias appraisal and does not support pooled effect estimates.

Conclusions

Clinical reasoning in health professions education is already supported by mature theories and broad assessment literature. AI has not replaced those foundations. It has exposed their misalignment. Current AI studies most often operationalise clinical reasoning through the parts that are easiest to prompt, simulate, document, or score. These studies are useful, but they do not yet tell the field how clinical reasoning should be taught, assessed, or protected when AI is present.

The next step is therefore framework building and testing. Such a framework as the one proposed should specify which dimensions of reasoning are being targeted, which pedagogical activities are expected to develop them, which assessment methods can produce defensible evidence, and which reasoning tasks should remain human-first at different stages of training. Without that alignment, the field will continue to mistake scalable proxies for the construct itself.

Declarations

Funding:

Nil

Ethics Declaration:

Not Applicable

Author Contribution:

NCW was responsible for the entirety of this manuscript.

References

1. Abdulnour, R. E. E., Gin, B. C., & Boscardin, C. K. (2025). Educational Strategies for Clinical Supervision of Artificial Intelligence Use. *The New England Journal of Medicine*, 393 8, 786–797. <https://doi.org/10.1056/nejmra2503232>
2. Abouzeid, E., & Sallam, M. A. (2025). AI-assisted script concordance tests: Enhancing feasibility with customized ChatGPT. *Medical Teacher*, 48, 757–760. <https://doi.org/10.1080/0142159X.2025.2533405>
3. Brügge, E., Ricchizzi, S., Arenbeck, M., Keller, M. N., Schur, L., Stummer, W., Holling, M., Lu, M., & 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. <https://doi.org/10.1186/s12909-024-06399-7>
4. Charlin, B., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. V. D. (2000). The Script Concordance Test: A Tool to Assess the Reflective Clinician. *Teaching and Learning in Medicine*, 12, 189–195. https://doi.org/10.1207/S15328015TLM1204_5
5. Charlin, B., Tardif, J., & Boshuizen, H. (2000). Scripts and Medical Diagnostic Knowledge: Theory and Applications for Clinical Reasoning Instruction and Research. *Academic Medicine*, 75, 182–190. <https://doi.org/10.1097/00001888-200002000-00020>
6. Cianciolo, A., LaVoie, N., & Parker, J. D. A. (2021). Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence. *Academic Medicine*, 96, 1026–1035. <https://doi.org/10.1097/ACM.0000000000004010>
7. Çiçek, F. E., Ülker, M., Özer, M., & Kiyak, Y. S. (2024). ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. *Postgraduate Medical Journal*. <https://doi.org/10.1093/postmj/qgae170>
8. Cooper, N., Bartlett, M., Gay, S., Hammond, A., Lillicrap, M., Matthan, J., & Singh, M. (2020). Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. *Medical Teacher*, 43, 152–159. <https://doi.org/10.1080/0142159x.2020.1842343>
9. Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine: Journal of the Association of American Medical Colleges*, 84 8, 1022–1028. <https://doi.org/10.1097/ACM.0b013e3181ace703>
10. Dai, X., Gong, Y., & Ma, L. (2026). Effects of ChatGPT-generated immediate feedback integrated into VR-based OSCEs on nursing students' performance: a randomized crossover study. *BMC Nursing*. <https://doi.org/10.1186/s12912-026-04633-9>
11. Daniel, M., Rencic, J., Durning, S., Holmboe, E., Santen, S., Lang, V. J., Ratcliffe, T. A., Gordon, D., Heist, B. S., Lubarsky, S., Estrada, C., Ballard, T., Artino, A., Silva, A. S. D., Cleary, T., Stojan, J. N., & Gruppen,

- L. (2019). Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. *Academic Medicine*. <https://doi.org/10.1097/ACM.0000000000002618>
12. Durning, S., Artino, A., Schuwirth, L., & van der Vleuten, C. V. D. (2013). Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. *Academic Medicine: Journal of the Association of American Medical Colleges*, 88(4), 442–448. <https://doi.org/10.1097/ACM.0b013e3182851b5b>
 13. Durning, S., Jung, E., Kim, D. H., & Lee, Y. M. (2024). Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside. *Korean Journal of Medical Education*, 36, 145–155. <https://doi.org/10.3946/kjme.2024.292>
 14. Goh, E., Gallo, R. J., 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). Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. *medRxiv*. <https://doi.org/10.1101/2024.03.12.24303785>
 15. Gordon, D., Rencic, J., Lang, V. J., Thomas, A., Young, M., & Durning, S. (2022). Advancing the assessment of clinical reasoning across the health professions: Definitional and methodologic recommendations. *Perspectives on Medical Education*, 11, 108–114. <https://doi.org/10.1007/s40037-022-00701-3>
 16. Guth, T. A., Wolfe, R., Martinez, O., Subhiyah, R. G., Henderik, J. J., McAllister, C., & Roussel, D. (2024). Assessment of Clinical Reasoning in Undergraduate Medical Education: A Pragmatic Approach to Programmatic Assessment. *Academic Medicine*, 99, 912–921. <https://doi.org/10.1097/ACM.0000000000005665>
 17. Hege, I., Kononowicz, A., Berman, N. B., Lenzer, B., & Kiesewetter, J. (2018). Advancing clinical reasoning in virtual patients – development and application of a conceptual framework. *GMS Journal for Medical Education*, 35. <https://doi.org/10.3205/zma001159>
 18. Holderried, F., Stegemann-Philipps, C., Herrmann-Werner, A., Festl-Wietek, T., Holderried, M., Eickhoff, C., & Mahling, M. (2024). A Language Model–Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR Medical Education*, 10. <https://doi.org/10.2196/59213>
 19. Hrynchak, P., Takahashi, S. G., & Nayer, M. (2014). Key-feature questions for assessment of clinical reasoning: a literature review. *Medical Education*, 48, 870–883. <https://doi.org/10.1111/medu.12509>
 20. Huang, H. M. (2026). Effectiveness of ChatGPT using the START-Chat model in enhancing nursing students' pediatric knowledge, clinical reasoning, and self-directed learning: A quasi-experimental study. *Clinical Simulation in Nursing*. <https://doi.org/10.1016/j.ecns.2026.101958>
 21. Hudon, A., Phan, V., Charlin, B., & Wittmer, R. (2025). Teaching Clinical Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests: Mixed Methods Formative Evaluation. *JMIR Formative Research*, 9. <https://doi.org/10.2196/76618>
 22. Kiyak, Y. S., Emekli, E., Kara, T. İ., Coşkun, Ö., & Budakoğlu, I. (2025). AI Teaches Surgical Diagnostic Reasoning to Medical Students: Evidence from an Experiment Using a Fully Automated, Low-Cost

- Feedback System. *Journal of Surgical Education*, 82(10), 103639.
<https://doi.org/10.1016/j.jsurg.2025.103639>
23. Lubarsky, S., Dory, V., Audétat, M., Custers, E., & Charlin, B. (2015). Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Canadian Medical Education Journal*, 6, e61–e70. <https://doi.org/10.36834/cmej.36631>
24. Lubarsky, S., Dory, V., Duggan, P., Gagnon, R., & Charlin, B. (2013). Script concordance testing: From theory to practice: AMEE Guide 75. *Medical Teacher*, 35, 184–193.
<https://doi.org/10.3109/0142159X.2013.760036>
25. Masters, K., MacNeil, H., Benjamin, J., Carver, T., Nemethy, K., Valanci-Aroesty, S., Taylor, D. C. M., Thoma, B., & Thesen, T. (2025). Artificial Intelligence in Health Professions Education assessment: AMEE Guide 178. *Medical Teacher*, 47, 1410–1424.
<https://doi.org/10.1080/0142159X.2024.2445037>
26. Nissen, L., Rother, J. F., Heinemann, M., Reimer, L., Jonas, S. M., & Raupach, T. (2025). A randomised cross-over trial assessing the impact of AI-generated individual feedback on written online assignments for medical students. *Medical Teacher*, 47, 1544–1550.
<https://doi.org/10.1080/0142159X.2025.2451870>
27. Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine*, 70, 194–201.
<https://doi.org/10.1097/00001888-199503000-00009>
28. Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: the dual-process theory. *Medical Education Online*, 16.
<https://doi.org/10.3402/meo.v16i0.5890>
29. Peterson, B. D., Magee, C. D., Martindale, J., Dreicer, J. J., Mutter, M. K., Young, G., Sacco, M., Parsons, L., Collins, S., Warburton, K. M., & Parsons, A. (2022). REACT: Rapid Evaluation Assessment of Clinical Reasoning Tool. *Journal of General Internal Medicine*, 37, 2224–2229.
<https://doi.org/10.1007/s11606-022-07513-5>
30. Plackett, R., Kassianos, A., Mylan, S., Kambouri, M., Raine, R., & Sheringham, J. (2022). The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review. *BMC Medical Education*, 22. <https://doi.org/10.1186/s12909-022-03410-x>
31. Rencic, J., Schuwirth, L., Gruppen, L., & Durning, S. (2020a). A situated cognition model for clinical reasoning performance assessment: a narrative review. *Diagnosis*, 7, 227–240.
<https://doi.org/10.1515/dx-2019-0106>
32. Rencic, J., Schuwirth, L., Gruppen, L., & Durning, S. (2020b). Clinical reasoning performance assessment: using situated cognition theory as a conceptual framework. *Diagnosis*, 7, 241–249.
<https://doi.org/10.1515/dx-2019-0051>
33. Rodrigues-Pereira, P., Dias-Calças, M. A. P., Mélo, A. M., Melchior, M., Ribeiro, L. G., Pazin-Filho, A., Mazzi-Chaves, J., & Magri, L. (2025). Generative Artificial Intelligence-Driven Clinical Case Simulation

- in Temporomandibular Disorder Education: ChatGPT Versus Real Patients. *Journal of Dental Education*. <https://doi.org/10.1002/jdd.70104>
34. 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. <https://doi.org/10.1186/s12909-026-09079-w>
 35. Saloojee, H., Gramanie, M. C., Mwali, R., Bassett, B. A., Madhi, S., & Kala, I. S. (2026). AI at the bedside: Randomised controlled trial of ChatGPT's impact on student performance in real-patient clinical exams. *Medical Teacher*. <https://doi.org/10.1080/0142159x.2026.2652061>
 36. Schaye, V., DiTullio, D., Guzman, B., Vennemeyer, S., Shih, H., Reinstein, I., Weber, D. E., Goodman, A., Wu, D. T. Y., Sartori, D. J., Santen, S. A., Gruppen, L. D., Aphinyanaphongs, Y., & Burk-Rafel, J. (2024). Large Language Model–Based Assessment of Clinical Reasoning Documentation in the Electronic Health Record Across Two Institutions: Development and Validation Study. *Journal of Medical Internet Research*, 27. <https://doi.org/10.2196/67967>
 37. Schaye, V., Guzman, B., Burk-Rafel, J., Marin, M., Reinstein, I., Kudlowitz, D., Miller, L., Chun, J. W., & Aphinyanaphongs, Y. (2022). Development and Validation of a Machine Learning Model for Automated Assessment of Resident Clinical Reasoning Documentation. *Journal of General Internal Medicine*, 37, 2230–2238. <https://doi.org/10.1007/s11606-022-07526-0>
 38. Schaye, V., Miller, L., Kudlowitz, D., Chun, J. W., Burk-Rafel, J., Cocks, P., Guzman, B., Aphinyanaphongs, Y., & Marin, M. (2021). Development of a Clinical Reasoning Documentation Assessment Tool for Resident and Fellow Admission Notes: a Shared Mental Model for Feedback. *Journal of General Internal Medicine*, 37, 507–512. <https://doi.org/10.1007/s11606-021-06805-6>
 39. Schuwirth, L., Durning, S., & King, S. M. (2020). Assessment of clinical reasoning: three evolutions of thought. *Diagnosis*, 7, 191–196. <https://doi.org/10.1515/dx-2019-0096>
 40. Sigirtmac, I. C., Turkmen, C., Değerli, Y. İ., & Oksuz, C. (2026). Effect of AI chatbot–assisted case-based learning on clinical reasoning in occupational therapy students: a post-test only randomized controlled trial in Parkinson's disease education. *Interactive Learning Environments*. <https://doi.org/10.1080/10494820.2026.2617984>
 41. Sur, M., Dissaneevate, S., Yuenyongviwat, A., Ruangnapa, K., Liaw, W., Boedeker, P., Kumar, S., Rismiller, B., & Thammasitboon, S. (2026). Leveraging artificial intelligence to accelerate competency development in diagnostic reasoning of medical students. *Medical Teacher*, 1–9. <https://doi.org/10.1080/0142159X.2026.2641769>
 42. Thammasitboon, S., Rencic, J., Trowbridge, R., Olson, A., Sur, M., & Dhaliwal, G. (2018). The Assessment of Reasoning Tool (ART): structuring the conversation between teachers and learners. *Diagnosis*, 5, 197–203. <https://doi.org/10.1515/dx-2018-0052>
 43. Thampy, H., Willert, E., & Ramani, S. (2019). Assessing Clinical Reasoning: Targeting the Higher Levels of the Pyramid. *Journal of General Internal Medicine*, 34, 1631–1636. <https://doi.org/10.1007/s11606-019-04953-4>

44. Torre, D., Daniel, M., Ratcliffe, T. A., Durning, S., Holmboe, E., & Schuwirth, L. (2024). Programmatic Assessment of Clinical Reasoning: New Opportunities to Meet an Ongoing Challenge. *Teaching and Learning in Medicine*, 37, 403–411. <https://doi.org/10.1080/10401334.2024.2333921>
45. Waechter, J., Kusnoor, A., Eickman, K., Smith, D., Wong, M., Rogers, S., & Lee, C. H. (2026). Reducing misdiagnoses and cognitive errors using virtual patients and automated feedback in a clinical reasoning curriculum. *BMC Medical Education*, 26. <https://doi.org/10.1186/s12909-026-08647-4>
46. Wu, Y. (2025). Future-proofing medical assessment: evaluating collaborative intelligence in the age of artificial intelligence. *Academic Medicine: Journal of the Association of American Medical Colleges*. <https://doi.org/10.1093/acamed/wvaf039>
47. Zafar, I., Farooq, M., Caliskan, S. A., & Magzoub, M. (2025). Ten Tips for AI-Assisted Key Feature Problems: A Validity-Informed Guide for Medical Education. *F1000Research*. <https://doi.org/10.12688/f1000research.173421.1>

Figures

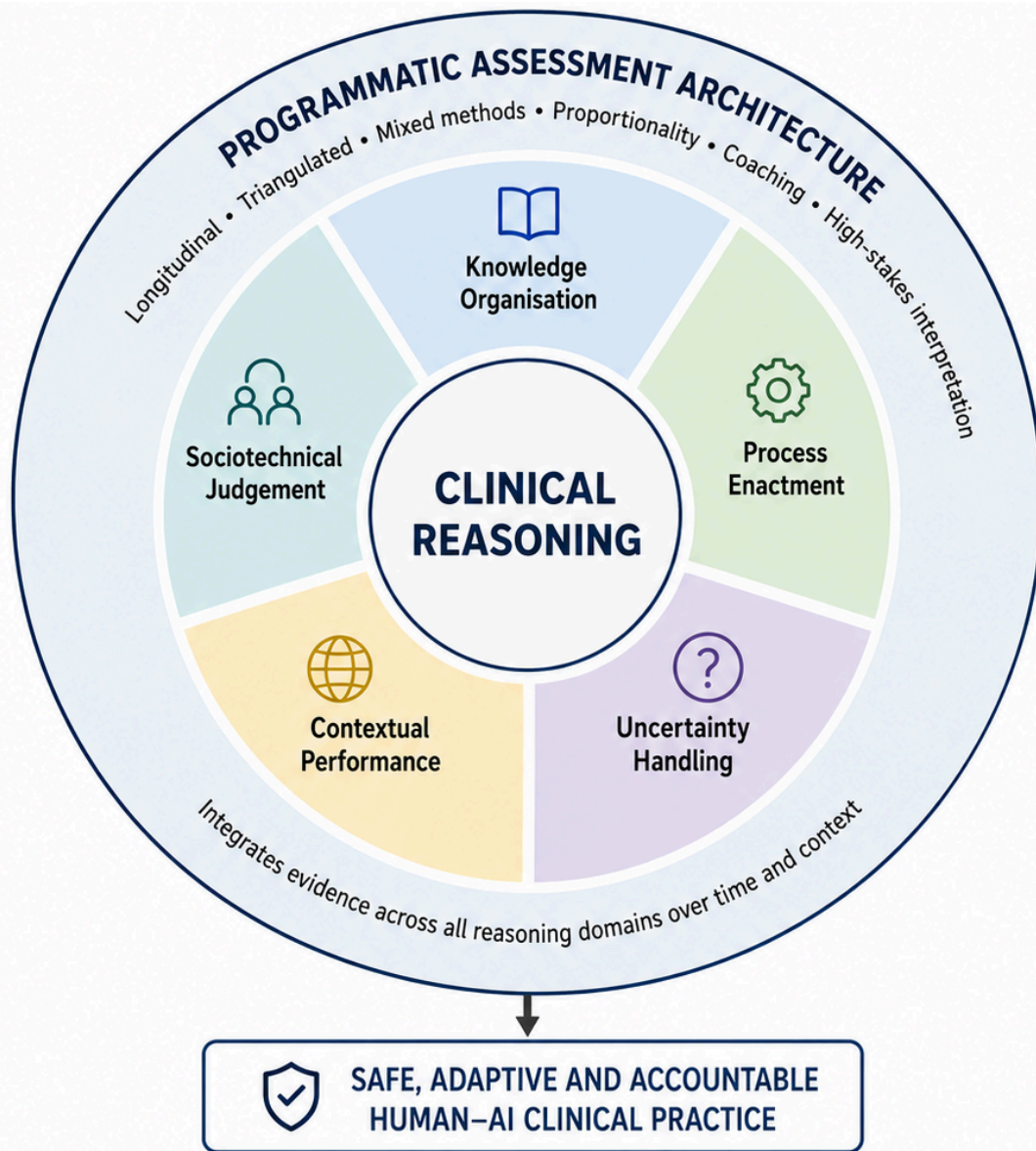


Figure 1. A multidimensional framework for clinical reasoning in health professions education in the era of artificial intelligence.

Clinical reasoning is conceptualised as a multidimensional construct comprising five interrelated domains: knowledge organisation, process enactment, uncertainty handling, contextual performance, and sociotechnical judgement. The outer layer, programmatic assessment architecture, reflects the need for longitudinal, triangulated, and contextually grounded assessment strategies that integrate evidence across domains rather than relying on isolated performance measures. Together, these components support the development and evaluation of safe, adaptive, and accountable human-AI clinical practice.

Figure 1

See image above for figure legend.