Cureus
Part of SPRINGER NATURE

Open Access Review Article

Generative Artificial Intelligence in Clinical Reasoning Education: A Focused Review and Framework

Review began 08/19/2026
Review ended 10/27/2026
Published 11/02/2026

© **Copyright** 2026
Qin et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.118991

Boquan Qin [1], Bingyu Wu [2], Peize Liang [3], Xiang Fang [1]

1. Department of Orthopedics, West China Hospital of Sichuan University, Chengdu, CHN 2. West China School of Medicine, West China Hospital of Sichuan University, Chengdu, CHN 3. Department of Ophthalmology, West China Hospital of Sichuan University, Chengdu, CHN

Corresponding author: Xiang Fang, fangxiang@wchscu.edu.cn

Abstract

Clinical reasoning education requires learners to generate and justify diagnostic plans under uncertainty, yet supervised practice and feedback are uneven. Generative artificial intelligence (AI), large language models, and virtual patients may expand practice, but improved assisted performance may not represent durable unaided reasoning. We conducted a focused narrative review of literature identified through PubMed, Scopus, Web of Science, and Google Scholar, supplemented by reference-list review and citation searching. The synthesis covered AI-powered virtual patients, large language model-based diagnostic coaching, and automated assessment or feedback, with emphasis on evidence involving learners and on indirect evidence that clarifies educational mechanisms or risks. Most studies report associations or short-term improvement rather than AI-specific educational effects, and several rely on satisfaction outcomes, pre-post designs, or historical controls. Current evidence therefore supports feasibility and selected short-term gains in practice, diagnostic performance, and feedback but does not establish durable learning. However, model errors, unreliable references, weaker performance on image-based tasks, and limited evidence for retention or transfer constrain high-stakes use. Recent reviews describe broad applications of AI in medical education, but less attention has been paid to when assistance should enter a reasoning task and how to preserve foundational competence. We propose AI-augmented cognitive apprenticeship, a framework organized around four principles: struggle first, AI second; AI for formative practice, humans for summative judgment; transparent AI literacy; and competence-calibrated AI assistance. The framework treats never-skilling, defined as failure to acquire a foundational competence, as a testable educational risk; operationally, never-skilling is failure to reach a predefined unaided competency threshold after AI-supported training. AI should therefore be introduced after an independent attempt, calibrated to demonstrated competence and task risk, and paired with periodic unaided assessment. Prospective longitudinal studies are needed to determine whether AI-supported practice transfers to durable clinical reasoning.

Categories: Therapeutics, Medical Education, Orthopedics
Keywords: artificial intelligence, clinical reasoning, cognitive apprenticeship, formative assessment, generative artificial intelligence, medical education, virtual patients

Introduction And Background

Clinical reasoning, the process by which clinicians gather and interpret information to reach diagnoses and management decisions, is central to medical competence [1-2]. Drawing on dual-process theory, it operates through an interplay of rapid, intuitive type 1 processes and deliberate, analytical type 2 processes, with experts fluidly moving between the two [1-2]. In surgical practice, clinical reasoning is uniquely demanding because it integrates diagnostic formulation, procedural risk stratification, and intraoperative decision-making, decisions that carry immediate and often irreversible consequences [1,3].

Clinical reasoning remains difficult to teach and assess. Cognitive apprenticeship, in which experts make their reasoning visible through modeling, coaching, scaffolding, and guided reflection, requires sustained faculty participation [4]. Bedside exposure varies with caseload and service structure, while objective structured clinical examinations (OSCEs) provide episodic samples of performance. Faculty ratings add contextual judgment but are also shaped by social and rater effects [5]. These limitations matter because diagnostic errors occur in an estimated 10-15% of clinical encounters [6], and surgical reasoning errors can affect decisions from preoperative interpretation through postoperative management [3,6].

The public release of ChatGPT in November 2022 marked an inflection point in the accessibility of generative artificial intelligence (AI) [7]. Within months, large language models (LLMs), that is, AI systems trained on large text corpora that generate text in response to a prompt, such as GPT-4, Claude, and DeepSeek, were available to trainees at all levels, and reports of their use proliferated across informal educational networks. In parallel, three kinds of structured educational tools have emerged: LLM-driven virtual patient (VP) platforms, meaning simulated clinical encounters in which a model plays the patient, capable of dynamic, unscripted clinical dialogue; natural language processing (NLP), the computational

analysis of free text, in systems that analyze student reasoning transcripts; and computer vision algorithms that provide automated feedback on procedural performance.

This rapid uptake has outpaced evidence about how AI should be integrated into learning. The Association of American Medical Colleges has articulated principles for responsible AI use in medical education [8]. However, broad principles do not specify when assistance should enter a reasoning task, which outcomes should be assessed without AI, or when human judgment must remain decisive. These questions are especially important in surgical training, where decisions are time-sensitive and often depend on imaging, procedural judgment, and longitudinal consequences.

We define AI tools for clinical reasoning education as systems that are designed or evaluated for teaching, that assess or provide feedback on a learner's reasoning process, or that provide indirect evidence that materially informs educational use. The third category includes clinical or examination studies only when they clarify a capability or risk relevant to learners. We exclude diagnostic systems evaluated solely for standalone clinical accuracy and performance benchmarks that do not inform an educational argument.

The review addresses AI-powered virtual patients for diagnostic practice, LLM-based diagnostic coaching, automated assessment of clinical reasoning, and the cross-cutting tension among cognitive offloading, deskilling, never-skilling, and upskilling. We use never-skilling to mean failure to acquire a foundational competence at all, as distinct from deskilling, the loss of a competence already acquired. We emphasize evidence relevant to surgical and orthopedic training while retaining studies from other specialties when they illuminate a transferable educational mechanism. Existing reviews describe AI applications in medical or surgical education broadly [9-13]. Comparatively less attention has been given to the distinction between AI-supported performance during training and durable, unaided reasoning after support is removed. That distinction organizes the present synthesis.

Review

Searches were executed on July 10, 2026, with a targeted update search on July 31, 2026, in PubMed, Scopus, Web of Science, and Google Scholar, for literature published primarily from January 2020 through June 2026; the update search extended coverage to July 31, 2026. The search concepts were organized into three blocks: AI, generative AI, large language models, ChatGPT, conversational AI, natural language processing, or AI-powered virtual patients; clinical reasoning, diagnostic reasoning, clinical decision-making, cognitive apprenticeship, cognitive offloading, deskilling, never-skilling, or upskilling; and medical education, health professions education, surgical education, orthopedic education, simulation, or virtual patients. Database-specific syntax was adapted to each platform, and the complete PubMed strategy is reproduced in Appendix A. Reference-list review and citation searching identified additional relevant publications.

We prioritized empirical studies of AI tools used for clinical reasoning education. Reviews, conceptual sources, and technical or clinical studies were included when they clarified an educational mechanism, capability, or risk. Indirect evidence was retained when it addressed model reliability, error patterns, or human-AI performance that could alter educational design, but such studies were not treated as evidence of learner benefit. Source selection was guided by relevance to clinical reasoning, educational function, study design, directness to learners, and applicability to surgical or orthopedic training. Purely administrative applications, procedural assessment without a reasoning component, and standalone diagnostic systems without a material educational implication were outside the review scope. English-language literature was prioritized, with targeted inclusion of relevant Chinese-language orthopedic studies. Publications before 2020 were retained only when they provided foundational theory or educational context.

We used thematic narrative synthesis to organize heterogeneous evidence by the educational function of AI: practice through virtual patients, diagnostic coaching, and assessment or feedback. Given the heterogeneity of study designs and evidence roles, we did not apply a single risk-of-bias or evidence-grading framework. No quantitative pooling was undertaken: no meta-analysis or meta-regression was performed, and no summary effect estimate is reported. Statistics are quoted from individual sources using the source's own metric, and effect sizes and 95% confidence intervals are reproduced wherever the source reports them. The synthesis keeps direct learner evidence distinct from performance benchmarks, indirect clinical evidence, and conceptual sources, while focusing on mechanisms and educational decisions. Records were screened independently by two authors (B.Q. and B.W.), first by title and abstract and then in full text; disagreements were resolved by discussion and by X.F. where agreement could not be reached. Because this is a narrative synthesis with a representative evidence base, per-stage record counts were not maintained and are not reported, and no per-study risk-of-bias score was assigned.

To allow readers to judge how directly each finding applies to surgical training, evidence is labelled in four tiers throughout. Tier 1 is direct learner evidence from surgical or orthopedic education, in which participants were surgical trainees and the outcome was a learner outcome. Tier 2 is direct learner evidence from other specialties or from undergraduate medical education, which informs educational mechanism but requires transfer to be established. Tier 3 is indirect evidence, comprising model-capability benchmarks and clinician-performance studies that describe what a system can do rather than what a learner gains; it is never treated as evidence of learner benefit. Tier 4 is conceptual or theoretical work that supplies the

educational mechanism being tested.

Recent Cureus reviews have synthesized broad applications of AI in medical education, professionalism, and surgical training [9-13]. We use these sources to position the field but focus more narrowly on the timing and function of AI assistance during clinical reasoning development. Our contribution is a competence-calibrated framework that distinguishes assisted performance from unaided competence and treats never-skilling as a testable educational risk.

AI-powered virtual patients: scaling deliberate diagnostic practice

The Evolution from Scripted to LLM-Driven Virtual Patients

The concept of virtual patients, computer-based simulations of clinical encounters designed for educational purposes, predates the current AI era by decades. However, pre-AI virtual patients were constrained by scripted branching logic, which required authors to anticipate every possible clinical trajectory. This increased development costs, limited realism, and prevented the systems from handling unexpected trainee questions or deviations from expected diagnostic pathways [14].

LLM-powered virtual patients extend this approach by supporting less scripted interactions. By leveraging the generative capabilities and broad medical knowledge embedded in models such as GPT-4, these systems can engage in unscripted, dynamically responsive clinical dialogues. A trainee can ask questions in natural language, and the virtual patient can respond with clinically coherent answers. Responses may express uncertainty or emotion and describe symptoms that vary as the simulated illness progresses [15-16].

Learner reception has generally been favorable. Liu et al. reported that approximately 84% of respondents were very satisfied or satisfied with the Intelligent Inquiry Training System (IITS) platform in a surgical curriculum [17], and Fąferek et al. found that ChatGPT-generated VPs were comparable with human-authored cases on selected ratings [16]. In a randomized study of 56 learners, GPT-simulated history-taking encounters produced higher post-training examination scores than role-play (86.79 versus 73.64, $p < 0.001$; Cohen's $d = 2.57$) [18].

Core Educational Evidence: Diagnostic Accuracy Improvement

Waechter et al. [19] conducted a multisite observational study across five North American medical schools. The analysis included 12,400 VP cases completed by 1,066 students and 1.55 million recorded data points. Learners completed up to 20 cases with automated scoring of diagnostic accuracy, justification, test ordering, and cognitive errors. The design permits analysis of performance trajectories but does not isolate the effect of the VP curriculum from repeated exposure, concurrent learning, or case-sequence effects.

Misdiagnosis occurred in 20.1% of cases. Within the observed sequence, the misdiagnosis rate declined from 27% on a learner's first case to 9.0% by the twentieth case ($p < 0.0001$). Diagnostic justification scores increased by 72%, test ordering improved by 40%, and recorded cognitive errors were halved. No plateau was observed over 20 cases. Together, these measures show progressive improvement across repeated VP practice during the 20-case sequence.

Diagnostic justification quality and cognitive error rate were the strongest independent predictors of diagnostic accuracy in multivariable analysis. This association supports attention to reasoning process rather than answer correctness alone. Performance among first-year students with greater case exposure further suggests that structured VP practice can engage learners early in clinical training.

Surgical Virtual Patients: OrthoSim and Beyond

The OrthoSim AI Agent, developed at Sun Yat-sen Memorial Hospital, Sun Yat-sen University, is a comparatively well-described surgical VP system [20]. It uses a multibranch, multi-outcome model to represent decisions from initial diagnosis through perioperative management and follow-up across four orthopedic subspecialties. Its distinctive feature is temporal consequence modeling: an earlier decision can alter the subsequent simulated course. This design makes consequences visible within the simulation, although the fidelity of those consequences and their transfer to clinical practice require further validation.

The single-center prospective study used historical controls. A total of 20 orthopedic graduate students received four weeks of supplementary OrthoSim training in addition to standard rotations, while 18 historical controls completed conventional rotations. The intervention group scored higher from the fourth Mini-Clinical Evaluation Exercise (Mini-CEX) onward and achieved a higher complication-management OSCE score (86.95 ± 9.97 versus 80.15 ± 4.04, $p = 0.019$). These results support feasibility and an association with short-term performance. These design features leave the independent contribution of OrthoSim uncertain.

Additional systems indicate feasibility across surgical contexts. A multi-agent neurosurgical platform

converted 608 self-assessment questions into conversational cases; residents outperformed GPT-4o in free-response diagnostic conversations (70.0% versus 28.3%) [21]. Rao et al. evaluated an LLM-based standardized-patient platform for cultural dexterity with 16 general surgery residents, who rated realism at 4.00/5 and reported value for reflective communication practice [22]. IITS, built on Baidu's ERNIE 4.0, also reported reduced faculty case-preparation workload [17]. Together, these studies show that AI-mediated cases can be implemented across multiple surgical learning contexts and can reduce selected faculty preparation demands.

Key Gaps in AI Virtual-Patient Research

The next evidentiary step for AI VPs is comparative and longitudinal evaluation. Existing studies are dominated by single-institution implementations, short follow-up, satisfaction outcomes, and pre-post or historical comparisons. Direct comparisons with time-matched human-precepted case discussion are lacking, as are measures of retention and unaided transfer. These comparisons would clarify when AI-mediated, expert-led, and blended practice offer distinct educational value.

LLM-based diagnostic coaching: evidence from educational benchmarks and clinical reasoning studies

Mechanisms and Theoretical Rationale

LLMs could support diagnostic coaching through several interaction designs. A Socratic design can ask learners to generate and prioritize hypotheses before revealing suggestions. An error-analysis design can compare a learner's explanation with alternative interpretations and prompt consideration of anchoring, premature closure, or confirmation bias. A calibration design can vary case difficulty or the amount of prompting. These are plausible educational functions, but their effectiveness depends on prompt design, source grounding, feedback timing, and the learner's ability to evaluate the output.

Interactivity distinguishes LLM coaching from static references, but interactivity alone is not deliberate practice. Deliberate practice requires a defined task, effortful performance, informative feedback, and repeated opportunities to improve [23]. An LLM may support some of these conditions, yet it may also reveal answers too early, provide incorrect feedback, or reward superficial fluency. Educational evaluation must therefore measure what the learner can subsequently do without assistance.

LLM Performance on Standardized Orthopedic Examinations

The Orthopaedic In-Training Examination (OITE) is a widely used benchmark in orthopedic training, and resident performance is associated with subsequent American Board of Orthopaedic Surgery (ABOS) Part I outcomes [24]. LLM performance on OITE questions therefore helps define capability and failure modes relevant to study-tool use. Educational value depends on how those capabilities are integrated into learning and feedback.

LLM performance on orthopedic examinations has increased across successive model generations. In 2023, ChatGPT (GPT-3.5) correctly answered approximately 47% of ABOS-style questions, roughly equivalent to a first-year resident and well below the passing threshold [25]. By late 2023, GPT-4 had improved to 74% on the 2022 OITE, exceeding the passing threshold and performing at a level statistically indistinguishable from postgraduate year 3 through postgraduate year 5 (PGY-3 to PGY-5) residents [26].

By 2025-2026, reasoning-optimized models achieved higher text-only benchmark scores. Diniz et al. [27] evaluated models on 702 questions; o3 achieved 93.6% accuracy and a 34% relative error reduction compared with GPT-4. Hayes et al. [28] tested 11 models on 119 text-based questions from the 2024 OITE. A total of 10 exceeded the ABOS minimum passing standard, eight exceeded mean postgraduate year 5 (PGY-5) performance, and o1Pro with Deep Research achieved 90.8%. These time-stamped evaluations show rapid gains in text-based orthopedic knowledge performance and provide benchmarks for educational testing.

The Reliability Gap: Accuracy Is Not Sufficient

High benchmark accuracy coexists with unreliable reasoning and referencing. Singh and Chandrasekhar administered 203 OITE questions to GPT-5 [29]. The model achieved 78.3% accuracy, but detailed review of 88 responses found fabricated or misrepresented references in 33% overall and 15.9% of correct answers. The subsample design limits population-wide estimates, yet the result shows why answer accuracy cannot serve as the sole criterion for an educational tool.

Jain et al. compared ChatGPT-4o responses on 635 OITE questions with official answer explanations [30]. Responses were classified as ideal in 58.7%, inadequate in 6.9%, and unacceptable in 34.4%. Accuracy was lower on image-based than text-only questions (60.0% versus 73.1%, $p < 0.001$), and unacceptable responses increased to 44.1% for media-based items. These results identify media-rich orthopedic questions as an

important failure condition for that model and examination context.

For educators, the relevant distinction is between solving an examination item and supporting a learner's development. Some models exceed mean senior-resident performance on selected text-only benchmarks, but plausible explanations and citations remain unreliable. Learners with limited knowledge may be least able to detect those errors. This pattern favors supervised use that makes model critique part of the learning task and measures later, unaided reasoning. First, this is indirect evidence: item-level accuracy on text-only questions establishes neither safe clinical tutoring, nor competence in image interpretation, nor suitability for summative assessment, because none of these functions is what the benchmark measures. Second, the figures are comparable only within the conditions each author team reported. Model version, prompting strategy, access to images or media, and test administration date were reported inconsistently across studies, and several reports describe a single administration of a single item set. We therefore report model, item set, and publication year for every result in this section and treat the reported accuracies as time-stamped capability estimates rather than as stable properties of a named system.

What AI Reasoning Patterns Reveal for Education

ClinDiag-GPT provides indirect evidence about human-AI diagnostic collaboration rather than an educational intervention [31]. The model was fine-tuned on 7,616 clinical cases and evaluated on 4,421 cases across 32 specialties using dynamic, multiturn diagnostic dialogues. Its design more closely resembles progressive information gathering than static question answering, but the participants and outcomes were not selected to measure learner development.

ClinDiag-GPT showed lower anchoring and confirmation-bias error rates than comparator models. In a three-arm comparison, physician-AI collaboration achieved 45.0% diagnostic accuracy, compared with 40.0% for physicians alone and 23.3% for AI alone [31]. This study supports the value of collaboration in that clinical setting; its implications for novice learners and over-reliance require direct educational evaluation.

A Chinese orthopedic study adds a second example of tool appraisal. DeepSeek R1 was evaluated on 87 preoperative cases, and junior physicians rated its clinical reference value highly for risk stratification [32]. These ratings inform perceived clinical usefulness and motivate educational studies that include learner outcomes and independent post-use performance.

Implications for Educational Practice

Taken together, the available studies support supervised educational testing of LLMs, in which the learner reasons first and AI is used to challenge or compare that reasoning before human-guided reflection; the framework proposed below converts this sequence into testable principles.

Automated assessment of clinical reasoning

The Assessment Gap

Clinical reasoning assessment in surgical training commonly combines structured encounters such as OSCEs with workplace-based faculty judgments. Structured encounters sample only a limited number of cases, while faculty judgments provide context but can be influenced by rater and social effects [5]. Neither approach routinely produces continuous information about how a learner gathered, prioritized, and integrated evidence.

AI could make reasoning processes more visible by analyzing questions, hypotheses, and revisions generated during a VP encounter. Such analysis may provide rapid, structured formative feedback; the output reflects the cases, scoring rules, reference standard, model version, and prompts. These dependencies must be examined before AI-generated scores are used for consequential decisions.

Evidence for Grading Performance

Schiött et al. evaluated LLM grading of clinical reasoning in rheumatology VP dialogues [33]. The best-performing model, o3-mini, achieved 81.5% grading correctness, compared with 82.3% for pooled human expert raters, and completed assessments more quickly. This near-parity concerns correctness in one task and dataset; it is not a reliability coefficient. Its result supports further validation of LLM-assisted grading across cases, learner groups, and surgical settings.

Chen et al. developed a coding scheme from 1,030 GPT-3.5 VP dialogues involving 210 students and five chest-pain cases [34]. A total of 12 behaviors across clinical reasoning, information gathering, and social interaction achieved interrater reliability of $\kappa \geqslant 0.85$. Summarizing, integrating, and logical organization correlated with several performance measures, whereas information-gathering quantity was less predictive of diagnostic accuracy. This work shows that process-oriented features can be defined and coded, but cross-case and cross-specialty generalization remains to be established.

AI-Assisted Faculty Feedback

AI may also support faculty feedback rather than score learners directly. Furey et al. used a generative writing tool to revise fixed-mindset statements; blinded coders classified 17 of 18 AI-modified statements (94.4%) as growth-mindset language [35]. This proof of concept shows that AI can help faculty transform feedback language and provides a basis for studying learner response and development.

In a surgical foundations program, fine-tuned GPT-3.5 achieved F1 scores of 0.827-0.949 across three Quality of Assessment for Learning (QuAL) dimensions when evaluating 2,229 entrustable professional activity (EPA) comments [36]. Ahmad et al. reported 90% concordance with faculty ($\kappa = 0.94$) in an initial set of 60 feedback entries before applying ChatGPT-4o to a larger otolaryngology dataset [37]. These studies support narrowly defined quality-improvement applications, while autonomous judgments about resident competence remain unvalidated.

From Agreement to Validity

Agreement with human ratings is necessary but insufficient for educational validity. An AI assessment should represent the intended reasoning construct, generalize across cases and learner groups, remain stable enough for its purpose, and produce acceptable consequences. Fairness, auditability, model updates, and the possibility that learners adapt their language to the scorer also require evaluation. Until these conditions are examined, AI-generated reasoning scores are better suited to low-stakes feedback and quality improvement than to summative progression decisions.

The Orthopedic Assessment Gap

Orthopedic surgical training remains an open validation setting for AI reasoning assessment. The ABOS Surgical Skills Assessment Program provides a structured formative framework that could inform future studies [38]. Research should begin with formative, auditable use and evaluate reasoning constructs separately from procedural execution, imaging interpretation, and global judgments of competence.

The deskilling debate: cognitive offloading, never-skilling, and implications for surgical training

Cognitive Apprenticeship Under Threat

Cognitive apprenticeship describes how experts make thinking visible, coach performance, provide scaffolding, and gradually withdraw support as learners gain competence [4]. The model assumes that learners attempt meaningful cognitive work and can explain how they reached a conclusion. AI can extend coaching opportunities, but assistance that supplies a conclusion before the learner has formed one may alter the work being learned.

The central educational distinction is between assisted performance during acquisition and learning that transfers to future, unaided contexts [39]. A learner may perform well with AI while developing either stronger metacognitive monitoring or greater dependence on external suggestions. Which outcome occurs is likely to depend on when AI enters the task, what it reveals, how errors are discussed, and whether independent performance is repeatedly assessed.

Cognitive Offloading: Theoretical Basis and Preliminary Evidence

Cognitive offloading is the use of external resources to reduce internal demand. Sparrow et al. found that expected online access changed what people remembered and where they expected to retrieve information [40]. Offloading can preserve attention for higher-order decisions; the educational risk arises when the offloaded process is itself a target competence, and learners are not required to retrieve or reconstruct it. Evidence from adjacent domains, where AI assistance has been shown to shift effort from generation to evaluation during writing and coding [41], describes the same mechanism, but the educational stakes differ: in clinical reasoning, the offloaded process is itself the competence being certified. Within medicine, the same concern has been raised for AI-mediated cognitive offloading specifically [42].

The desirable-difficulties framework predicts that manageable challenge during acquisition can improve retention and transfer [39]. Requiring a learner to generate and defend a differential before viewing AI output is therefore a plausible design principle. It remains a hypothesis in AI-supported clinical reasoning and should be tested against alternative timing and feedback conditions.

Existing studies provide indirect grounds for testing the hypothesis. Waechter et al. paired repeated practice with feedback on the learner's own reasoning rather than delegating diagnosis to AI [19]. ClinDiag-GPT performed worse alone than physicians alone in its clinical comparison [31]. Together, these findings identify assistance timing and task allocation as variables for prospective studies of later unaided

performance.

Never-Skilling: A Concept Critical for Surgical Training

Oettl et al. distinguish deskilling, the loss of an established skill through disuse, from never-skilling, the failure to acquire a foundational competence [43]. The distinction is particularly relevant to training because the two problems require different outcomes and time scales. Deskilling requires evidence of decline from a prior level. Never-skilling requires evidence that a competence failed to develop under conditions in which it otherwise would have developed. Each requires evidence beyond assisted performance.

We propose the following operational definition: never-skilling is failure to reach a predefined unaided competency threshold by a specified time point, despite AI-supported training of documented dose, in learners who would otherwise be expected to reach that threshold. Three elements make the concept measurable. First, the threshold must be specified a priori on a validated assessment completed without AI, for example, a predefined diagnostic accuracy, a rating on an OSCE, or an entrustment level on a workplace-based assessment. Second, the dose and time course of AI exposure must be recorded. Third, a comparator is required: a concurrent cohort trained without the assistance, a historical cohort, or the learner's own pre-exposure trajectory. Longitudinal measurement therefore requires repeated unaided assessments at fixed intervals, retention and transfer tasks rather than repeated exposure to the same cases, and reporting of both the proportion of learners below threshold and the size of any shortfall. A decline from a threshold the learner previously met indicates deskilling; failure to reach it at all indicates never-skilling. The distinction is not new: anxiety that cognitive aids will deskill physicians has recurred with each generation of technology [44], which is one reason the present debate is better framed as an empirical question about outcomes than as a judgment about tools.

Image interpretation provides a concrete test case. In one OITE study, ChatGPT-4o accuracy was lower for image-based than text-only questions [30]. The mismatch highlights a practical concern: trainees may be least able to evaluate AI output in domains where they have not yet developed perceptual expertise. Future studies can compare how different timings of AI assistance affect later unaided interpretation.

The Counter-Argument: AI as Upskilling

The alternative is upskilling: well-timed AI use may expand case exposure, make reasoning visible, support feedback, and preserve faculty time for higher-order coaching [43]. Under this account, offloading lower-value operations could allow greater attention to synthesis, uncertainty, communication, or procedural planning. These benefits depend on reliable output, appropriate task allocation, and opportunities to practice without assistance. ClinDiag-GPT offers one clinical example of augmentation: assistance improved diagnostic accuracy in that setting, while whether redistributed cognitive effort produces stronger independent reasoning remains an empirical question [31].

An Unresolved Tension

Longitudinal learner evidence is the next requirement for resolving the deskilling-upskilling question. Studies should follow AI-exposed surgical cohorts into independent practice and use validated clinical-reasoning measures designed to detect cognitive offloading and never-skilling. The most informative designs would vary assistance conditions and measure unaided performance, retention, transfer, error detection, and calibration over time.

Interpretation and framework

Summary of Key Findings

Across the three domains, AI is most consistently supported as a means of extending practice and formative feedback. This support rests largely on associational and short-term evidence from heterogeneous designs, including satisfaction surveys, pre-post comparisons, and historical controls. AI-VP systems can deliver repeated practice at scale, and selected studies report favorable short-term trajectories or outcomes [18-20]. Advanced LLMs perform strongly on text-only orthopedic examinations, with important variation in explanation quality, reference reliability, and image-based performance [27-30]. AI grading and feedback tools also show promising agreement or classification performance in defined settings [33-37]. The evidence therefore favors supervised, task-specific augmentation while durable learning and summative use remain priority validation targets. Table *1* highlights sources that anchor the argument.

Source (year)	Context and relevance	Design/source type	Sample or evidence base	AI application	Principal finding (effect size or CI where reported)	Evidence tier	Key limitation
Liu et al.		Implementation	700 questionnaires;	IITS using	Approximately 84% of		Single-institution survey; satisfaction

(2026) [17]: IITS	Surgical curriculum	study	487 valid responses	Baidu ERNIE 4.0	respondents reported positive satisfaction	Tier 1	outcome only; no unaided assessment
Wang et al. (2025) [18]	General medical education	Randomized controlled trial	56 students	GPT-simulated patients	Examination scores favored GPT-simulated patients (86.8 vs 73.6)	Tier 2	Single center; 56 students; post-training examination outcome, no retention measure
Waechter et al. (2026) [19]	General medical education	Multisite observational	12,400 cases completed by 1,066 students	Virtual-patient platform	Misdiagnosis decreased from 27% to 9% over repeated cases	Tier 2	No control group; effect of repeated exposure not separable from curriculum
Xu et al. (2026) [20]: OrthoSim	Orthopedic education	Historical control	38 residents	OrthoSim AI Agent	OSCE scores were higher with OrthoSim (87.0 vs 80.2, p = 0.019)	Tier 1	Single center; historical controls; 38 residents; supplementary exposure
Lum (2023) [25]	Orthopedic performance benchmark with educational relevance	Cross-sectional	207 OITE questions	ChatGPT using GPT-3.5	ChatGPT achieved 47% accuracy, approximately a PGY-1 performance level.	Tier 3	One model version on one item set; text-only
Diniz et al. (2026) [27]	Orthopedic performance benchmark with educational relevance	Cross-sectional	702 questions	Multiple LLMs	o3 achieved 93.6% accuracy	Tier 3	Multiple-choice accuracy only; no reasoning process or learner outcome
Hayes et al. (2026) [28]	Orthopedic performance benchmark with educational relevance	Cross-sectional	119 text-based questions across 11 models	Advanced LLMs	o1Pro with Deep Research achieved 90.8% accuracy	Tier 3	Text-based items only; single examination year; no image items
Singh and Chandrasekhar (2026) [29]	Orthopedic performance benchmark with educational relevance	Cross-sectional	203 OITE questions	GPT-5	Reference problems were found in 33% of audited responses	Tier 3	Reference audit on a subsample of responses; single model
Jain et al. (2025) [30]	Orthopedic performance benchmark with educational relevance	Cross-sectional	635 OITE questions	ChatGPT-4o	Unacceptable responses occurred in 34.4% of answers	Tier 3	One model; adds explanation quality and media items, but no learner outcome
Chen et al. (2026) [31]: ClinDiag-GPT	Multispecialty clinical diagnosis with indirect educational relevance	Multi-arm clinical evaluation	Model trained on 7,616 cases and benchmarked on 4,421 cases	ClinDiag-GPT	Physician-AI collaboration outperformed physicians alone in the reported comparison	Tier 3	Clinician-AI collaboration study; participants not selected to measure learning
Schiött et al. (2025) [33]	Rheumatology virtual patients	Comparative	Virtual-patient dialogues	o3-mini	o3-mini achieved 81.5% grading correctness, close to pooled human raters	Tier 2	One specialty and dataset; grading correctness, not a reliability coefficient
Chen et al. (2026) [34]	General medical education	Validation	1,030 dialogues from 210 students	GPT-3.5 virtual patients with a coding scheme	The coding scheme achieved interrater reliability of $\kappa \geq 0.85$	Tier 2	Five chest-pain cases and one model; cross-case and cross-specialty transfer untested
Furey et al. (2025) [35]	General surgery feedback	Qualitative	83 feedback statements	Google AI	AI converted 17 of 18 (94.4%) AI-modified feedback statements to growth-mindset language	Tier 2	Proof of concept on 18 statements; no learner-level outcome
Kwan et al. (2025) [36]	Surgical foundations feedback	Retrospective	2,229 EPA comments	Fine-tuned GPT-3.5	F1 scores ranged from 0.827 to 0.949 across QuAL	Tier 2	Retrospective; single program; comment-

					dimensions		level agreement only
Ahmad et al. (2025) [37]	Otolaryngology feedback	Retrospective	776 feedback entries	ChatGPT-4o	Initial concordance with faculty ratings was 90%	Tier 2	Retrospective; single specialty; agreement with faculty is not validity evidence
Oettl et al. (2026) [43]	Orthopedics and medical education	Narrative review	Literature synthesis without an original dataset	AI-assisted clinical workflows	The review distinguishes deskilling, never-skilling, and upskilling	Tier 4	Narrative review; no original dataset

TABLE 1: Selected empirical studies and conceptual sources informing artificial intelligence and clinical reasoning education

ABOS: American Board of Orthopaedic Surgery; EPA: entrustable professional activity; IITS: Intelligent Inquiry Training System; LLM: large language model; OITE: Orthopaedic In-Training Examination; OSCE: Objective Structured Clinical Examination; PGY: postgraduate year; QuAL: quality of assessment for learning

Evidence tiers: Tier 1, direct learner evidence from surgical or orthopedic education; Tier 2, direct learner evidence from other specialties or undergraduate medical education; Tier 3, indirect evidence from model-capability benchmarks and clinician-performance studies, which describe what a system can do rather than what a learner gains; Tier 4, conceptual or theoretical sources. Effect sizes and confidence intervals are reproduced where the source reports them; where a source reports only a p-value or a descriptive comparison, the source's own metric is given

Mechanisms and Boundary Conditions

Immediate performance and durable learning can change in opposite directions, so they require separate measurement. Immediate performance may improve because information is easier to retrieve or because the model supplies a missing hypothesis. Durable learning depends on whether the learner still retrieves, organizes, and tests knowledge when assistance is absent. Timing therefore matters: assistance after an independent attempt preserves a different learning task from assistance that supplies the first formulation.

Five mechanisms should be distinguished. Cognitive offloading is temporary delegation of mental work and may be productive or unproductive. Automation bias is inappropriate deference to a recommendation. Deskilling is decline in an established competence. Never-skilling is failure to acquire that competence. Upskilling describes improved capability through effective human-AI coordination. Each mechanism requires a different baseline, comparator, and outcome.

Likely boundary conditions include prior competence, task complexity, consequence severity, reversibility, feedback quality, model transparency, and availability of expert oversight. Assistance that is appropriate for a reversible practice case may be inappropriate during a high-stakes assessment. A useful framework must therefore calibrate support to demonstrated capability and task conditions rather than assume that chronological training stage is an adequate proxy.

Proposed Framework: AI-Augmented Cognitive Apprenticeship

Figure *1* summarizes a competence-calibrated implementation of AI-augmented cognitive apprenticeship for curriculum design and prospective testing. Assistance is calibrated to demonstrated competence, task risk, and the learner's ability to detect AI error. Postgraduate year may provide local context, but demonstrated competence and task conditions determine the form and timing of support.

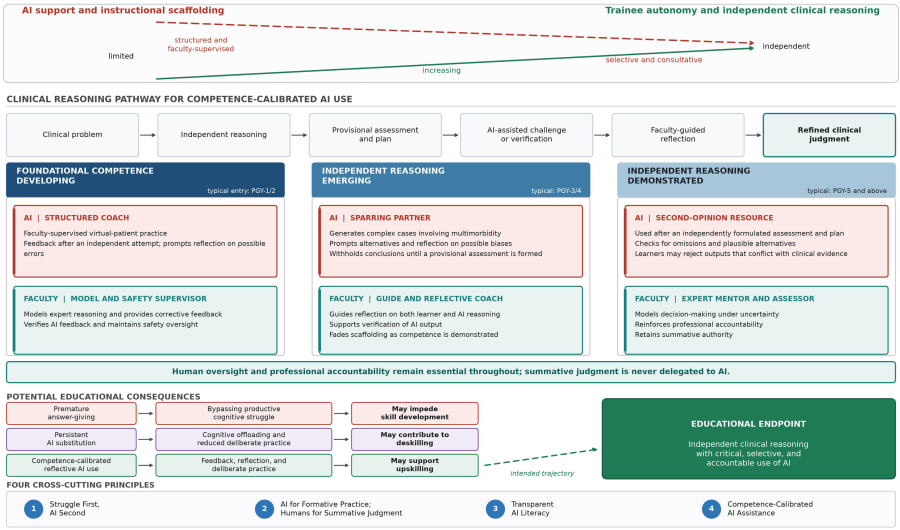


FIGURE 1: Competence-calibrated implementation of AI-augmented cognitive apprenticeship

AI: artificial intelligence

Top: as trainees move from limited to independent clinical reasoning, AI support becomes selective and consultative rather than structured and faculty-supervised, while trainee autonomy increases. Middle: the reason-first learning cycle, in which the learner reasons independently and commits to a provisional assessment and plan before AI-assisted challenge or verification, followed by faculty-guided reflection. Lower: three competence bands (foundational competence developing; independent reasoning emerging; independent reasoning demonstrated) with the corresponding roles of AI and of faculty; the typical postgraduate year is shown for orientation only and does not determine the band. Bottom: potential educational consequences of premature answer-giving, persistent AI substitution, and competence-calibrated reflective use, with human oversight and professional accountability retained throughout. The four cross-cutting principles are struggle first, AI second; AI for formative practice, humans for summative judgment; transparent AI literacy; and competence-calibrated AI assistance. These pathways are conceptual hypotheses, not established causal effects or quantitative thresholds. The figure was created by the authors using Python 3

The framework retains expert modeling, articulation, reflection, and accountability while using AI to extend selected coaching and practice functions. Its four principles combine evidence-supported safeguards with theory-derived design hypotheses, and each is linked to outcomes that can be tested prospectively.

Principle 1

Struggle first, AI second: Learners first generate and justify a provisional assessment before consulting AI. The principle is grounded in desirable difficulties and cognitive apprenticeship [4,39]. Outcomes include retention and unaided transfer alongside immediate assisted accuracy.

Principle 2

AI for formative practice, humans for summative judgment: AI can provide repeated cases and rapid, structured feedback, whereas faculty retain responsibility for modeling expert judgment, resolving disagreement, and making consequential competence decisions. Evidence from formative applications supports repeated practice and structured feedback [19,33-37], while faculty retain summative authority.

Principle 3

Transparent AI literacy as a curricular competency: Learners should understand model versioning, hallucination, reference failure, uncertainty, bias, privacy, and the limits of benchmark performance. The need is supported by documented failures in orthopedic question answering [29-30].

Principle 4

Competence-calibrated AI assistance: The amount and timing of assistance should depend on demonstrated unaided performance, task risk, reversibility, and error-detection ability. Learners who cannot yet formulate or critique a plan may need structured, supervised use after an independent attempt. More competent learners may use AI as a sparring partner or selective second opinion.

Worked example: An orthopedic imaging case. A second-year resident is shown the lateral radiograph of a distal radius fracture with dorsal comminution. Before any AI use, the resident writes the fracture pattern, states whether intra-articular extension is present, proposes computed tomography if step-off is suspected, and records the reasoning for each step. The resident then asks a large language model to review the same image; the model proposes a different classification and omits the step-off criterion. In faculty debrief, the resident must decide which output to accept, name the model's omission, and specify what evidence would settle the disagreement. The cycle closes with an unaided read of a different image two weeks later, so that what is measured is the resident's gain rather than the model's output. In the earlier band, the same sequence runs with faculty supervising the AI interaction directly; in the later band, the resident uses the model as a second opinion and carries the obligation to justify any disagreement.

Implications for clinical education and surgical training

Programs should avoid both unrestricted adoption and blanket prohibition. A staged pilot can define the target competence, specify when AI becomes available, require documentation of the learner's pre-AI reasoning, and include faculty-guided comparison of learner and model output. Each AI-supported activity should be paired with periodic unaided assessment so that improved tool-mediated performance is not mistaken for independent learning.

Orthopedic imaging illustrates the need for task-specific calibration. ChatGPT-4o performed less well on image-based OITE items in one study [30], while image interpretation remains a core training outcome. Early learners should therefore complete and document an independent interpretation before receiving AI suggestions. OrthoSim provides an operational example of AI-VP integration in orthopedic training [20]; pairing such systems with faculty debriefing and periodic unaided assessment would extend evaluation from implementation to educational effect.

The framework is not tied to a specialty or to a technology stack. Although the examples here are drawn largely from orthopedic and surgical training, the principles apply wherever clinical reasoning is taught and assessed, and the resource requirement is deliberately minimal: a web-accessible model, a written record of the learner's pre-AI reasoning, and a scheduled unaided reassessment. What is harder to fund is the faculty time for debrief and governance, and programs should cost that component before adoption rather than assume that generative AI removes it.

Implementation also requires faculty development and governance. Educators need sufficient AI literacy to detect overreliance, interpret model disagreement, and lead reflection. Programs should address privacy, data retention, bias, access inequities, model updates, and the additional workload of supervising AI-generated feedback [8]. These conditions may determine whether artificial intelligence expands learning opportunities or shifts educational burden without improving learning.

Limitations

The field changes rapidly, so model-specific findings are time-stamped rather than stable capability estimates. Many educational studies are small, single-center, and short-term. Assessment evidence is largely non-orthopedic, and transfer to surgical reasoning is uncertain. Targeted Chinese-language retrieval did not address other non-English literature. The proposed framework is therefore a theory-informed synthesis that requires prospective evaluation. Because the search was structured but not exhaustive and per-stage screening counts were not maintained, relevant sources cannot be excluded, and the synthesis is offered as representative rather than exhaustive.

Future research priorities

Longitudinal studies should document the timing, frequency, and function of AI exposure and repeatedly assess unaided reasoning. Cluster-randomized, stepped-wedge, encouragement, and carefully controlled cohort designs could compare competence-calibrated AI use with alternative assistance schedules while reducing contamination and ethical concerns. Primary outcomes should include unaided diagnostic performance, retention, and transfer; secondary outcomes should include error detection, calibration, cognitive-bias patterns, resource use, and learner autonomy. Trials comparing AI-VP practice with time-matched expert-precepted discussion would clarify comparative effectiveness. Instrument development is also needed to distinguish productive offloading, automation bias, deskilling, and never-skilling. Because model versions change quickly, studies should report model, prompt, grounding sources, update date, and failure conditions. Multisite curricula with shared competency outcomes would improve comparability.

Conclusions

AI can expand practice and feedback opportunities in clinical reasoning education, with selected studies showing favorable short-term outcomes across VP practice, orthopedic knowledge tasks, and automated feedback. Across these domains, the decisive question is whether supported performance develops into durable, unaided reasoning. Never-skilling is therefore best treated as a testable training outcome alongside retention, transfer, error detection, and calibration. AI-augmented cognitive apprenticeship translates this question into a practical design for prospective evaluation. It places independent reasoning before AI

consultation, reserves summative judgment for humans, treats AI literacy as a curriculum outcome, and calibrates assistance to demonstrated competence and task risk. Programs can pilot these principles with faculty oversight and repeated assessment of unaided reasoning, generating the longitudinal evidence needed to refine training-stage recommendations.

Appendices

Appendix A. PubMed search strategy

("Artificial Intelligence"[MeSH Terms] OR "artificial intelligence"[tiab] OR "generative artificial intelligence"[tiab] OR "large language model*"[tiab] OR ChatGPT[tiab] OR "conversational artificial intelligence"[tiab] OR "natural language processing"[tiab] OR "virtual patient*"[tiab]) AND ("clinical reasoning"[tiab] OR "diagnostic reasoning"[tiab] OR "clinical decision-making"[tiab] OR "cognitive apprenticeship"[tiab] OR "cognitive offloading"[tiab] OR deskilling[tiab] OR "never-skilling"[tiab] OR upskilling[tiab]) AND ("medical education"[tiab] OR "health professions education"[tiab] OR "surgical education"[tiab] OR "orthopedic education"[tiab] OR simulation[tiab] OR "virtual patient*"[tiab]) AND ("2020/01/01"[PDAT] : "2026/07/31"[PDAT])

The same three concept blocks were adapted to the syntax of Scopus and Web of Science; Google Scholar was used for citation searching only. The strategy was last executed on July 31, 2026.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Xiang Fang, Boquan Qin

Drafting of the manuscript: Xiang Fang, Boquan Qin

Critical review of the manuscript for important intellectual content: Xiang Fang, Peize Liang, Bingyu Wu, Boquan Qin

Supervision: Xiang Fang

Acquisition, analysis, or interpretation of data: Peize Liang, Bingyu Wu

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

ChatGPT 5.6 (OpenAI) was used during manuscript preparation to assist with language editing, organization, and improving the clarity of the text. All AI-assisted content was critically reviewed and revised by the authors, who take full responsibility for the accuracy, integrity, and originality of the final manuscript.

References

1. Croskerry P: A universal model of diagnostic reasoning . Acad Med. 2009, 84:1022-8. 10.1097/ACM.0b013e3181ace703
2. Kahneman D: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York; 2011.
3. Loftus TJ, Tighe PJ, Filiberto AC, et al.: Artificial intelligence and surgical decision-making. JAMA Surg. 2020, 155:148-58. 10.1001/jamasurg.2019.4917
4. Collins A, Brown JS, Holum A: Cognitive apprenticeship: making thinking visible . Am Educ. 1991, 15:6-11, 38-46.
5. Gingerich A, Regehr G, Eva KW: Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 2011, 86:S1-7. 10.1097/ACM.0b013e31822a6cf8
6. Graber ML: The incidence of diagnostic error in medicine . BMJ Qual Saf. 2013, 22:ii21-7. 10.1136/bmjqs-2012-001615
7. Boscardin CK, Gin B, Golde PB, Hauer KE: ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. 2024, 99:22-7. 10.1097/ACM.0000000000005439
8. Principles for the responsible use of artificial intelligence in and for medical education . (2025). Accessed: August 10, 2026: https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use .

9. Hallquist E, Gupta I, Montalbano M, Loukas M: Applications of artificial intelligence in medical education: a systematic review. Cureus. 2025, 17:e79878. 10.7759/cureus.79878
10. Komasawa N, Yokohira M: Generative artificial intelligence (AI) in medical education: a narrative review of the challenges and possibilities for future professionalism. Cureus. 2025, 17:e86316. 10.7759/cureus.86316
11. Almansour M, Soliman M, Aldekhyyel R, Binkheder S, Temsah MH, Malki KH: An academic viewpoint (2025) on the integration of generative artificial intelligence in medical education: transforming learning and practices. Cureus. 2025, 17:e81145. 10.7759/cureus.81145
12. Al-Saadawi A, Tehranchi S, Ahmed S, Nzeako OJ: Exploring the current applications of artificial intelligence in orthopaedic surgical training: a systematic scoping review. Cureus. 2025, 17:e81671. 10.7759/cureus.81671
13. Zerrouki Y, Baran JV, Bandi R, Moothedan E, Knecht MK, Follin T, Perumareddi P: Artificial intelligence in U.S. surgical training: a scoping review mapping current applications and identifying gaps for future research applications. Cureus. 2025, 17:e93793. 10.7759/cureus.93793
14. Cook DA, Triola MM: Virtual patients: a critical literature review and proposed next steps . Med Educ. 2009, 43:303-11. 10.1111/j.1365-2923.2008.03286.x
15. Yu H, Zhou J, Li L, et al.: Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. Commun Med (Lond). 2025, 6:27. 10.1038/s43856-025-01283-x
16. Fąferek J, Kononowicz AA, Bogutska N, et al.: Applying ChatGPT to plan and create a realistic collection of virtual patients for clinical reasoning training. BMC Med Educ. 2025, 25:1277. 10.1186/s12909-025-08006-9
17. Liu H, Cheng H, Zhang Y, et al.: Revolutionizing clinical skills training with generative AI: the Intelligent Inquiry Training System framework. Front Med (Lausanne). 2026, 13:1781314. 10.3389/fmed.2026.1781314
18. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC: Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. BMC Med Educ. 2025, 25:1030. 10.1186/s12909-025-07614-9
19. Waechter J, Kusnoor A, Eickman K, Smith D, Wong M, Rogers S, Lee CH: Reducing misdiagnoses and cognitive errors using virtual patients and automated feedback in a clinical reasoning curriculum. BMC Med Educ. 2026, 26:414. 10.1186/s12909-026-08647-4
20. Xu J, Li D, Lu D, et al.: Application and evaluation of a virtual patient system based on a multi-branch dynamic decision model in orthopedic graduate medical education. BMC Med Educ. 2026, 26:956. 10.1186/s12909-026-09327-z
21. Sangwon KL, Zhang J, Steele R, et al.: A multi-AI agent framework for interactive neurosurgical education and evaluation: from vignettes to virtual conversations. Neurosurg Pract. 2026, 7:e000217. 10.1227/neuprac.0000000000000217
22. Rao AS, Lee RS, Bott E, et al.: The digital standardized patient: an artificial intelligence coach for cultural dexterity in surgical care. J Am Coll Surg. 2025, 241:887-94. 10.1097/XCS.0000000000001505
23. Ericsson KA: Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med. 2004, 79:S70-81. 10.1097/00001888-200410001-00022
24. Swanson D, Marsh JL, Hurwitz S, et al.: Utility of AAOS OITE scores in predicting ABOS part I outcomes: AAOS exhibit selection. J Bone Joint Surg Am. 2013, 95:e84. 10.2106/JBJS.L.00457
25. Lum ZC: Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023, 481:1623-30. 10.1097/CORR.0000000000002704
26. Xu AY, Singh M, Balmaceno-Criss M, et al.: Comparitive performance of artificial intelligence-based large language models on the orthopedic in-training examination. J Orthop Surg (Hong Kong). 2025, 33:10.1177/10225536241268789
27. Diniz P, Yokoe T, Öttl FC, Pereira H, Henriques R, Samuelsson K: Reasoning-optimised large language models reach near-expert accuracy on board-style orthopaedic exams: a multi-model comparison on 702 multiple-choice questions. Knee Surg Sports Traumatol Arthrosc. 2026, 34:752-62. 10.1002/ksa.70222
28. Hayes DS, Barré A, Muchow RD: Assessing the third wave of generative AI: performance of advanced models on text-based questions from the 2024 Orthopaedic In-training Examination. J Am Acad Orthop Surg. 2026, 34:e1636-48. 10.5435/JAAOS-D-25-00441
29. Singh S, Chandrasekhar P: Accuracy is not enough: reasoning and reference reliability in orthopaedic large language model (LLM) applications. Cureus. 2026, 18:e100845. 10.7759/cureus.100845
30. Jain N, Gottlich C, Fisher J, Winston T, Matullo K, Greenhill D: ChatGPT-4o is not a reliable study source for orthopaedic surgery residents. JB JS Open Access. 2025, 10: 10.2106/JBJS.OA.25.00112
31. Chen X, Zhou H, Yi H, et al.: Grounding large language models in clinical diagnostics. Nat Commun. 2026, 17:4401. 10.1038/s41467-026-70274-w
32. Wei SS, Wu HB, Li SL, Wen ZL, Yang CA, Lu QS, Liu PL: Decision-making performance and auxiliary value of large language models in preoperative management of orthopedic surgery. J Shandong Univ (Health Sci). 2026, 64:104-10.
33. Schiött J, Ivegren W, Borg A, Parodis I, Skantze G: Using LLMs to grade clinical reasoning for medical students in virtual patient dialogues. Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2025). 2025, 750-63.
34. Chen N, Tang L, Liu Y, et al.: Developing and validating a coding scheme for clinical reasoning in history taking using generative AI-based virtual patients: systematic text condensation approach. JMIR Med Educ. 2026, 12:e84347. 10.2196/84347
35. Furey MJ, Stemrich R, Westfall-Snyder J, Gupta T, Rapp M, Hoffman RL: Can artificial intelligence coach faculty to utilize growth mindset language? A qualitative analysis of feedback statements. J Surg Res. 2025, 308:300-6. 10.1016/j.jss.2025.01.029
36. Kwan BY, Zhou Z, Rogoza N, Aghaei N, de Vries I, Hanmore T, Zevin B: Leveraging large language models to evaluate the quality of narrative feedback for surgery residents in competency-based medical education. Ann Surg Open. 2025, 6:e608. 10.1097/AS9.0000000000000608
37. Ahmad SA, Armache M, Trakimas DR, Chen JX, Galaiya D: Evaluating resident feedback using a large language model: are we missing core competencies?. Laryngoscope. 2025, 135:4119-24. 10.1002/lary.32368

38. American Board of Orthopaedic Surgery: ABOS surgical skills assessment program . (2019). Accessed: August 10, 2026: https://www.abos.org/abosksb/.
39. Bjork RA: Memory and metamemory considerations in the training of human beings . Metacognition: Knowing About Knowing. Metcalfe J, Shimamura AP (ed): MIT Press, Cambridge, MA; 1994. 185-205. 10.7551/mitpress/4561.003.0011
40. Sparrow B, Liu J, Wegner DM: Google effects on memory: cognitive consequences of having information at our fingertips. Science. 2011, 333:776-8. 10.1126/science.1207745
41. Cash TN, Kelly MO, Macnamara BN, Risko EF: Is AI making us stupid?. Trends Cogn Sci. 2026, 30:673-5. 10.1016/j.tics.2026.06.004
42. Viola P, Chiarella G: Artificial intelligence and cognitive offloading: a potential new form of cognitive deprivation?. Eur Arch Otorhinolaryngol. 2026, 283:4765-6. 10.1007/s00405-026-10210-2
43. Oettl FC, Pruneski J, Zsidai B, et al.: From de-skilling to up-skilling: how artificial intelligence will augment the modern physician. J Exp Orthop. 2026, 13:e70677. 10.1002/jeo2.70677
44. Lea AS: Cognitive aids, artificial intelligence, and deskilling in medicine: the history of an enduring anxiety . NEJM AI. 2026, 3:AIp2500932. 10.1056/AIp2500932