

Review Article

Surgical training in the era of artificial intelligence and large language models: a narrative review from the perspective of a surgical resident

Ashwin Vinod^{1*}, Varun Venkatesh², George Mathew John³

¹Department of Urology, KLES Dr. Prabhakar Kore Hospital and Medical Research Centre, Jawaharlal Nehru Medical College, Kaher, Belagavi, Karnataka, India

²Department of Surgical Oncology, Regional Cancer Centre, Thiruvananthapuram, Kerala, India

³Department of General Surgery, Amrita Institute of Medical Sciences, Amrita Vishwa Vidyapeetham, Kochi, Kerala, India

Received: 27 August 2026

Accepted: 07 October 2026

*Correspondence:

Dr. Ashwin Vinod,

E-mail: ashwinvinod04@gmail.com

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Artificial intelligence (AI) and large language models (LLMs) have entered surgical residency faster than the curricula, assessment systems and governance frameworks designed to accommodate them, yet most published appraisals adopt the standpoint of educators rather than that of the trainee, who is simultaneously the principal user and the principal subject of these tools. This narrative review, informed by the SANRA criteria, searched five databases from January 2015 to June 2026 and synthesises the evidence from the resident's perspective. Genuine educational value is supported in four domains: on-demand knowledge synthesis and examination preparation, intelligent tutoring and automated feedback in simulation, objective video- and kinematics-based skill assessment, and relief from administrative burden. Randomised trials show that AI tutoring improves machine-derived performance metrics but that gains are often not reproduced on blinded human ratings, and that AI combined with expert instruction outperforms AI alone. Against this stand substantial risks: hallucination and guideline drift, with accuracy falling sharply on current, guideline-specific content; the first real-world signal of clinical deskilling following routine AI exposure in a procedural specialty; cognitive offloading, with attendant mis-skilling and never-skilling during formative training; threats to assessment construct validity and to the apprenticeship relationship; breaches of confidentiality; algorithmic bias; and a widening divide between well-resourced and resource-constrained programmes. Surveys show substantial trainee uptake alongside negligible formal instruction. The governing principle for the trainee is sequence: independent reasoning and unassisted practice must precede augmentation. Training programmes should teach structured AI literacy, protect AI-free deliberate practice, and adopt verification-centred supervision.

Keywords: Artificial intelligence, ChatGPT, Deskillling, Large language models, Residency training, Surgical education

INTRODUCTION

Surgical training has always been an apprenticeship. The Halstedian model, for all its subsequent modification, rests on a simple transaction: a senior surgeon lends judgement and hands to a junior one, repeatedly, until the junior can be trusted to work alone. The apparatus that

has grown around this transaction over three decades has refined it without replacing it. The apparatus is now under pressure from several directions. Working-hour restrictions and expanding non-clinical responsibilities have compressed operative exposure; one widely cited analysis suggested that a substantial minority of graduating surgical residents did not meet independent

competence in core procedures.¹ Case complexity has increased, minimally invasive and robotic platforms have lengthened learning curves, and the volume of guideline-level knowledge a trainee is expected to command has grown beyond what can be committed to memory. Assessment remains largely dependent on the observation of a single supervising surgeon and is limited by rater bias, availability and the time cost of structured evaluation^{2,3}

Into this environment arrived, in late 2022, a class of tools that residents adopted almost immediately and largely without instruction. Generative artificial intelligence (AI) built on large language models (LLMs) can summarise a guideline chapter in seconds, generate a differential, draft a discharge summary, construct a viva-style question bank and rewrite a manuscript paragraph into fluent academic English. In parallel, but less visibly to the trainee, computer vision and machine learning applied to operative video and robotic kinematics began to make automated assessment of technical skill a practical proposition rather than a laboratory curiosity.⁴

Most published commentary on this transition is written from above. The trainee's position differs in kind, not merely in degree. The resident is at once the heaviest user of these tools and their principal subject: the person whose operative video will be scored by an algorithm, whose progression may one day be influenced by a machine-derived competence trajectory, and whose clinical reasoning is still being formed at precisely the moment that a fluent and confident machine offers to do the reasoning instead. This review takes that standpoint deliberately, surveying what the evidence supports, what it does not, and what a resident might reasonably conclude about using these tools without being diminished by them.

METHODS

This is a narrative review, prepared in accordance with the Scale for the Assessment of Narrative Review Articles (SANRA), which provides quality criteria for non-systematic syntheses where the aim is breadth of conceptual coverage rather than reproducible exhaustiveness.⁵

PubMed/MEDLINE, Scopus, Web of Science, Embase and Google Scholar were searched from 1 January 2015 to 30 June 2026. Terms were combined using Boolean operators and included: ("artificial intelligence" OR "machine learning" OR "deep learning" OR "computer vision" OR "large language model*" OR "ChatGPT" OR "generative artificial intelligence") AND ("surgical education" OR "surgical training" OR "residency" OR "resident*" OR "trainee*" OR "postgraduate medical education" OR "competency-based medical education"). Targeted searches additionally addressed automated skill assessment, intelligent tutoring systems, examination performance of LLMs in surgical specialties, deskilling

and publication ethics. Reference lists of retrieved articles were hand-searched.

No restriction was placed on surgical specialty, but the density of published evidence proved markedly uneven. Neurosurgical simulation and urological examination benchmarking furnish the most robust datasets, and gastrointestinal endoscopy provides the only real-world deskilling signal. Findings from these specialties are used throughout as exemplars, and the limits of extrapolation are addressed in the discussion. Eligible sources included randomised controlled trials, cohort and cross-sectional studies, systematic and scoping reviews, professional society position statements, and ethical guidance from medical journal organizations. Priority was given to primary studies in surgical or procedural specialties, to randomised evidence where available, and to publications from 2023 onwards given the pace of technological change. Preprints and conference abstracts were consulted for orientation but were not relied upon for substantive claims. No formal risk-of-bias assessment or quantitative synthesis was undertaken, consistent with the narrative design.

RESULTS

A working taxonomy for the trainee

"AI" is used loosely in surgical discourse, and the looseness matters because risks and benefits differ sharply between categories. Narrow, task-specific AI comprises supervised models trained to perform a single defined task on structured input, largely computer-vision systems applied to operative video and instrument kinematics: phase recognition, instrument detection and tracking, gesture recognition and skill classification.^{4,6} These do not converse, do not generalize beyond their training distribution, and fail silently rather than fluently.

Generative AI and LLMs are transformer-based models trained on very large text corpora to predict subsequent tokens. Their output is linguistically fluent by construction and factually correct only contingently. This is the single most important property for a trainee to internalise: fluency and accuracy are produced by different mechanisms, and the former is a poor proxy for the latter. Retrieval-augmented systems couple a generative model to a curated document store so that answers are grounded in, and traceable to, specified source material; this substantially reduces but does not abolish fabrication, and introduces a dependency on the currency of the underlying corpus.

Autonomous systems represent the frontier. Language-conditioned imitation learning has been used to execute a defined phase of a laparoscopic procedure, clipping and cutting during cholecystectomy, on ex vivo tissue without human intervention and with self-correction of intermediate errors.⁷ This is a proof of concept in a controlled setting rather than a clinical capability, but it

establishes that step-level autonomy is an engineering problem rather than a conceptual impossibility.

Knowledge acquisition and clinical reference

The most universal use is also the least glamorous. LLMs function as a tireless, instantly available tutor for factual and conceptual material, at a level of availability no human resource can match. For trainees in high-volume, service-heavy programmes this is not a trivial gain.

The performance data require careful reading and carry a lesson that generalizes beyond any one specialty. Urology is used here as the worked example because it offers the densest benchmarking against both board examinations and a mature guideline corpus. On standardised urological board-style items, GPT-4 answered 44.4% correctly against 30.9% for GPT-3.5, well below the 60% minimum passing standard proposed for the American Board of Urology's continuing certification activity.⁸ On the Taiwan Urology Board Examination, GPT-4 achieved 57.8% against 33.8% for GPT-3.5, with a declining accuracy trend across successive examination years and significantly lower accuracy in frequently updated domains (53.2%) than in stable domains (62.2%).⁹ Comparable instability has been documented on European Board of Urology in-service assessment items, where accuracy differed markedly between model generations and fluctuated between repeat testing rounds of the same model.¹⁰

The pattern recurs when accuracy is examined against guidelines rather than examination keys. When GPT-4 was asked eleven clinical decision-making questions on urolithiasis, every response contained guideline-concordant material, yet six omitted guideline-specified components and eight contained incorrect data.¹¹

Comparable model-dependent variability was found when four leading models were benchmarked against cardiothoracic surgery self-assessment items, indicating that this is not an artefact of one specialty's question style.¹² Performance degrades precisely where the trainee most needs precision: nuanced, current, guideline-specific, algorithmic content. A general surgery trainee querying the timing of source control in necrotising pancreatitis, an orthopaedic trainee querying implant selection thresholds and a urology trainee querying stone metaphylaxis are all interrogating the same vulnerable region of model competence. One development qualifies this pessimism substantially. A urology-specialised retrieval-augmented pipeline grounded in European Association of Urology guidelines achieved 88.4% correct answers on 200 board in-service assessment questions, against 77.6% for the underlying general model and a reported historical urologist average of 68.7%, while remaining clinician-verifiable because each assertion could be traced to a guideline passage.¹³ The practical implication is direct: a general-purpose chatbot is a study companion of uncertain reliability, whereas a guideline-grounded, source-citing system is a reference tool. These are not the same object and should not be trusted equally (Table 1).

Table 1: Reported performance of large language models on surgical board and guideline-based assessment items.

Setting and item bank	Model(s)	Reported accuracy	Key qualifier
US standardised urology board-style items (n=700)	GPT-3.5 vs GPT-4	30.9% vs 44.4%	Both below the 60% proposed minimum passing standard
Taiwan Urology Board Examination, 2020-2022 (n=450)	GPT-3.5 vs GPT-4	33.8% vs 57.8%	Declining accuracy by year; lower accuracy in frequently updated domains (53.2% vs 62.2%)
European Board of Urology in-service assessment (n=100)	GPT-3.5, GPT-4, Bing AI	58-62%, 63-77%, 73-81%	Variable across repeat testing rounds
European Board of Urology in-service assessment (n=200)	GPT-4o vs guideline-grounded retrieval-augmented pipeline	77.6% vs 88.4%	Historical urologist mean 68.7%; answers verifiable against guideline text
EAU-derived urolithiasis decision-making vignettes (n=11)	GPT-4	Not scored as % correct	6/11 answers omitted guideline components; 8/11 contained incorrect data
American Board of Thoracic Surgery SESATS XIII (n=400)	GPT-3.5, GPT-4, Med-PaLM 2, Claude 2	Model-dependent	Comparative analysis across four models

EAU: European Association of Urology; SESATS: Self-Education Self-Assessment in Thoracic Surgery. Urology is over-represented because it provides the densest published benchmarking against both board examinations and a mature guideline corpus; the pattern of degradation on current, guideline-specific content is reproduced in other specialties. Data from references 8-13.

Simulation and technical skill acquisition

The most methodologically rigorous evidence comes from randomised trials of AI tutoring in virtual reality simulation, and it is more interesting than the headlines suggest. In a multi-institutional randomised trial of 70 medical students performing simulated subpial brain tumour resections, participants randomised to an AI tutoring system built on an explainable machine-learning model of surgical expertise achieved higher machine-derived performance scores during practice and on a transfer task than those receiving remote expert instruction or no feedback.¹⁴ Crucially, blinded Objective Structured Assessment of Technical Skills (OSATS) ratings did not differ significantly between groups.¹⁵ That discrepancy is the finding, not a footnote: apparent benefit depended on whether performance was judged by the same system that delivered the training or by independent human raters. A secondary analysis demonstrated that competencies selected by the AI produced unintended changes in trainee behaviour that human instructors were needed to detect and correct.¹⁶

Later trials refined rather than reversed this picture. A randomised trial of 97 trainees comparing real-time

multifaceted AI feedback with in-person expert instruction found distinct advantages for each modality and a four-arm trial demonstrated that the modality of automated feedback itself materially alters learning outcomes.^{17,18} A 2025 randomised trial comparing AI-augmented personalised expert instruction against AI tutoring alone found the combination superior for performance and skill transfer.¹⁹ The consistent signal is that AI tutoring is a genuine additive intervention and a poor substitute. It scales feedback, standardises it and delivers it without fatigue, and it does not perceive what an experienced surgeon perceives. Operationally, AI tutoring can absorb the repetitive, metric-driven component of deliberate practice, reserving scarce faculty time for the judgement-laden and tacit components that machines cannot yet teach (Table 2).

Table 2: Randomised evidence on artificial intelligence tutoring in simulated surgical skills training.

Study	Design and participants	Comparison	Principal finding
Fazlollahi et al	Multi-institutional RCT; 70 medical students; VR subpial tumour resection	AI tutor vs remote expert instruction vs no feedback	AI group superior on machine-derived scores; blinded OSATS ratings did not differ between groups
Fazlollahi et al	Secondary analysis of the above RCT	AI-selected vs expert-selected competencies	AI-selected training targets produced unintended behavioural changes requiring human detection
Yilmaz et al	RCT; 97 trainees	Real-time multifaceted AI vs in-person expert instruction vs no real-time feedback	Distinct advantages for each modality; neither uniformly superior
Yilmaz et al	Four-arm RCT	Differing automated feedback modalities	Feedback modality materially affects learning outcomes
Giglio et al	Single-blinded RCT; medical students; VR simulator	AI-augmented personalised expert instruction vs AI tutoring alone	Combined instruction superior for performance, skill transfer and affective-cognitive response

AI: artificial intelligence; OSATS: Objective Structured Assessment of Technical Skills; RCT: randomised controlled trial; VR: virtual reality. Data from references 15-19.

Objective assessment of technical skill

Surgical assessment has a long-acknowledged validity problem: it is subjective, observer-dependent, episodic and expensive. Systematic reviews of automated technical skill assessment document models performing phase recognition, instrument tracking, gesture and action recognition and skill classification, with moderate to high accuracy in distinguishing novice from expert performance and reasonable agreement with expert raters.^{4,6,23,24}

Robotic platforms are especially amenable because they generate rich kinematic, sensor and video data as a by-

product of normal operation.^{20,21,22} For the trainee this offers continuous rather than episodic evaluation, longitudinal skill trajectories, feedback within minutes rather than weeks, and potential freedom from halo effects and inter-rater variability. A resident who can see that instrument-path efficiency has improved across thirty cases has something no logbook has provided. The field remains immature.

A systematic review of objective and AI-based tools in robotic skill assessment found extreme heterogeneity, few well-validated instruments and almost no deployment outside research settings, and called for collaboration between surgical societies, AI scientists and industry

around large, high-quality datasets and open benchmarks.²⁰

Administrative burden, feedback and scholarship

The benefit residents report most immediately is rarely educational in the formal sense. Discharge summaries, referral letters, operative notes, audit datasets, consent documentation and departmental reports consume hours that are, from a training standpoint, dead. LLMs draft these competently, and extraction and summarization of structured clinical text is among the more mature applications of the technology. Reduced documentation burden does not automatically become learning time, but it removes a barrier that no curricular reform has managed to remove. Similarly, competency-based training generates a volume of narrative feedback that is never systematically analysed because programmes lack the human reading time; triage and thematic synthesis of such free-text is precisely what language models do well. This remains a prospect rather than established practice and warrants prospective evaluation before any such output informs progression decisions.

For residents in academic programmes, particularly those for whom English is a second or third language, generative AI has lowered the barrier to scholarly output through literature triage, argument structuring, statistical code and language editing. The equity dimension is real and too often dismissed: reviewers have long conflated linguistic polish with scientific quality. The constraints are ethical rather than technical and are now well codified. The International Committee of Medical Journal Editors, the World Association of Medical Editors and the Committee on Publication Ethics converge: chatbots cannot be authors, because they cannot approve a final version, be accountable for integrity, or declare conflicts of interest; authors must disclose AI use at submission with tool and version; authors bear full responsibility for AI-generated content including accuracy and freedom from plagiarism; and reviewers must not upload manuscripts to systems that cannot guarantee confidentiality.²⁵⁻²⁸ Documented problems with hallucinated citations make independent verification of every reference non-negotiable.²⁹

Hallucination and the illusion of fluency

LLMs generate confident, well-structured, plausible text irrespective of whether the content is correct. For a consultant with two decades of pattern recognition an error is often conspicuous; for a resident it frequently is not. This asymmetry is the core epistemic hazard of the technology in training: the trainee lacks precisely the knowledge base required to audit the tool supplying the knowledge. Accuracy falls disproportionately on guideline-specific and recently updated content, the areas where a trainee is most likely to be questioned and least able to detect error.^{9,11} Models are poorly calibrated, conveying uniform confidence regardless of reliability.

Prompt sensitivity compounds this: in one board examination evaluation a simple prompt outperformed elaborate strategy-based prompting (60% versus 40%), so the trainee cannot rely even on prompt sophistication as a proxy for output quality.⁹

Cognitive offloading and the deskilling triad

This is the most serious objection, and in at least one procedural specialty it now rests on empirical evidence. In a multicentre observational study across four Polish endoscopy centres, the adenoma detection rate of standard, non-AI-assisted colonoscopy fell from 28.4% before the introduction of computer-aided detection to 22.4% afterwards, an absolute reduction of 6.0% and a relative reduction of approximately 20%, with AI exposure an independent predictor on multivariable analysis.^{30,31} These were experienced endoscopists, and the skill loss appeared not while the tool was in use but the moment it was withdrawn. This is the first real-world demonstration in medicine of automation-associated skill erosion linked to a patient-relevant quality indicator. If experienced operators deskill, the implications for trainees acquiring the skill in the first place are considerably more serious, and the study authors identified this as the priority question for further work.³⁰

Recent educational scholarship disaggregates the harm into three phenomena. Deskilling is erosion of an established competence in an experienced practitioner. Mis-skilling is uncritical internalization of AI-generated errors as fact, so that flawed knowledge becomes the trainee's own. Never-skilling, the concept most specific to residency, is failure to develop a foundational competence at all because the task was outsourced before the underlying capacity was ever built.³² A resident who has never constructed an unaided differential for painless haematuria, never planned an operative approach without prompting and never written an operative note from memory has not offloaded a skill; they have never possessed it. Direct evidence of these effects in surgical trainees is presently absent. The concern is grounded in learning theory and in the deskilling signal from an adjacent procedural specialty, and should be treated as a hypothesis warranting prospective study rather than an established finding.

Assessment, apprenticeship and equity

Automated skill assessment carries obvious appeal to programme administrators and a less obvious threat to trainees, and the threat is one of construct validity. Current models are trained predominantly on the measurable, such as instrument path length, economy of motion and phase duration, because these are what the data contain. The determinants that most reliably separate a safe surgeon from an unsafe one include recognizing when to stop, when to convert, when to call for help and when the anatomy does not match expectation. None is well represented in current metric sets. A system trained

on efficiency may reward a trainee who dissects rapidly through an unclear plane over one who pauses appropriately. Optimizing trainees against an incomplete metric is a known failure mode of measurement systems, and the finding that few tools have been validated outside research settings bears directly on any proposal to use them in progression decisions.²⁰

The trainee-mentor relationship is not merely a transmission channel for information; it carries professional identity formation, calibration of judgement and socialization into responsibility. Frictionless machine answers risk reducing the frequency of the effortful, formative conversation in which a trainee must expose the limits of their reasoning and be corrected. There is also an inversion of expertise: residents are frequently more fluent with these tools than their supervisors, so the person best placed to supervise their use often understands them least.³² Survey data underline the mismatch. In a cross-sectional survey of surgeons, trainees and students in Ireland, 72.3% reported no AI training despite 86.2% expressing interest, with accuracy and reliability (38.7%) and absence of evidence of effectiveness (33.7%) the leading concerns, and trainees expressing greater concern than consultants.³³ A large international survey in urology similarly documented rapid, largely self-directed uptake.³⁹ Trainees are self-teaching a high-stakes skill in the absence of a curriculum.

Pasting a clinical history into a consumer chatbot is a disclosure of personal health data to a third-party processor, frequently across an international border, often under terms permitting retention and model training. Most residents doing this have not read the terms and most institutions have issued no policy. The World Association of Medical Editors has warned that chatbots

retain the prompts fed to them, so that supplying a manuscript breaches confidentiality; the logic applies with greater force to patient data.²⁶ Regulatory frameworks are consolidating, including the European Union's AI Act, India's Digital Personal Data Protection Act, 2023, national research-ethics guidance and international guidance on large multimodal models, but they have not reached the level of ward-level rules a resident on a night shift can apply.³⁵⁻³⁸

Models also reproduce the distribution of their training data, which in surgery is drawn overwhelmingly from high-income, high-volume, well-instrumented centres. Clinically, recommendations may not reflect local disease patterns, presentation stage or resource availability. Educationally, a skill-assessment model trained on the operative style of one surgical culture may score a technically sound but stylistically different trainee as deficient. Documented sociodemographic biases in LLM medical decision-making make this a demonstrated rather than theoretical risk.³⁴

A resource asymmetry compounds it: validated, guideline-grounded, institutionally licensed systems cost money, while free consumer models are the least reliable and least secure. Where a well-resourced programme trains residents on a governed platform and a constrained programme's residents use an unlicensed chatbot on a personal phone, the technology has widened the very gap it was expected to close. Finally, when a trainee acts on an AI recommendation and harm results, responsibility rests with the trainee and the supervising surgeon; no framework transfers liability to a developer, and the human-in-the-loop construction can function as responsibility-shifting where that human has been deskilled by the system, they are supposed to check.^{30,32}

Table 3: Principal risks of artificial intelligence for the surgical trainee and corresponding mitigations.

Risk	Mechanism and evidence	Mitigation at trainee level	Mitigation at programme level
Hallucination and guideline drift	Fluency generated independently of accuracy; accuracy falls on guideline-specific and frequently updated content	Verify every actionable output against the primary guideline or trial	Provide access to retrieval-augmented, source-citing systems
Cognitive offloading and deskilling	Unassisted adenoma detection fell from 28.4% to 22.4% after routine AI exposure	Schedule deliberate AI-free practice; monitor unassisted performance	Protect AI-free assessment and training periods
Never-skilling and mis-skilling	Foundational competence never acquired, or AI errors internalised as fact	Build the skill before delegating the task	Stage AI introduction by training year and competence level
Reduction to metrics	Models trained on measurable efficiency rather than judgement; few tools validated outside research settings	Interrogate what a metric does and does not capture	Use automated assessment formatively; publish metrics; permit appeal
Confidentiality and	Prompts are retained by	De-identify absolutely	Provide a governed, contractually

Continued.

Risk	Mechanism and evidence	Mitigation at trainee level	Mitigation at programme level
data protection	third-party processors	before any prompt	compliant institutional deployment
Algorithmic bias and inequity	Training data skewed toward high-income, high-volume centres; sociodemographic bias documented in model outputs	Interpret recommendations against local epidemiology and resources	Validate tools locally; fund equitable access
Publication ethics	Fabricated references and undisclosed generative use	Verify every citation; disclose per ICMJE, WAME and COPE	Institutional policy and research-integrity training
Accountability	Liability remains with the clinician; human-in-the-loop may shift rather than share responsibility	Never act on output you cannot independently justify	Define supervisory and consent standards before deployment

AI: artificial intelligence; COPE: Committee on Publication Ethics; ICMJE: International Committee of Medical Journal Editors; WAME: World Association of Medical Editors.

DISCUSSION

The evidence supports neither abstention nor uncritical adoption, but a discipline of sequence and verification. Competence must precede augmentation: form your own plan first, then consult the model. Used as a second opinion an LLM strengthens reasoning by forcing articulation and comparison; used as a first opinion it substitutes for the cognitive work that constitutes learning.³² This ordering costs perhaps ninety seconds per encounter and is the single highest-yield behavioural rule available to a trainee. Because the deskilling signal implies that unassisted capability must be maintained by exercising it, trainees should preserve protected periods of practice performed entirely without AI assistance and should notice when unassisted performance begins to feel harder than it used to. Every clinically actionable output should be checked against the primary guideline or trial, every reference independently confirmed, and retrieval-augmented source-citing systems preferred over general chatbots for clinical content. Prompts should be treated as clinical data and de-identified absolutely. AI use should be disclosed in scholarly work without exception and discussed openly with supervisors, since concealment is what converts a legitimate tool into a professionalism problem.

Supervision should be engaged structurally rather than defensively. The DEFT-AI approach proposed for clinical supervision, which diagnoses the AI moment, explores the inputs supplied to the model, gives feedback on the learner's own reasoning independent of the tool, teaches verification and advises on future use, converts an AI interaction into a teaching encounter in under two minutes and requires no institutional infrastructure.²⁶ It is the most readily implementable intervention identified in this review, and trainees can initiate it as easily as supervisors.

For programmes, the survey evidence describes a consistent and correctable failure: widespread trainee use, minimal formal instruction, low confidence in appraisal and strong demand for curricular integration. Structured AI literacy should be introduced as a defined competency rather than an optional extra, covering functional model architecture, characteristic failure modes, prompt construction, verification technique, data-protection obligations and the evidence base for and against clinical use. Assessment requires redesign rather than defence: where written work is produced outside direct supervision its provenance can no longer be assumed, so programmes should grade process rather than product and strengthen observed, in-person assessment. Automated skill-assessment tools should be introduced formatively and transparently, with published metrics, demonstrated construct validity for judgement as well as efficiency and a right of appeal, and should not be used summatively until validated in the relevant population. Faculty development is the rate-limiting step, since supervisors cannot scaffold critical appraisal of a tool they have not used; educators should model sceptical usage explicitly by verbalising their own verification during rounds and operative teaching. Institutions should provide governed access, because leaving residents to use whichever consumer model is free is simultaneously a data-protection failure and an equity failure, and a licensed, de-identifying, guideline-grounded deployment with a written policy is achievable at modest cost.

Several limitations qualify these conclusions. This review is narrative rather than systematic: the search, while broad, was not exhaustive, no formal risk-of-bias appraisal or quantitative synthesis was performed, and selection of sources reflects the authors' judgement of relevance. Sources were restricted to peer-reviewed publications with fully verifiable bibliographic records, which excludes some relevant conference abstracts and preprints and may under-represent the most recent survey

literature. The evidence base is unevenly distributed across specialties, with randomised simulation data derived principally from neurosurgery, examination benchmarking from urology and the single real-world deskilling observation from gastrointestinal endoscopy; extrapolation across disciplines is reasoned rather than demonstrated and requires specialty-specific replication. Much published work in this field comprises small single-centre studies, evaluations with simulated rather than real learners, and outcomes measured immediately rather than longitudinally, with comparisons frequently made against the AI system's own metric rather than an independent standard. Models are superseded within months, so specific performance figures should be read as characteristic of a model generation rather than durable properties of the technology. Almost no study has followed trainees long enough to determine whether AI-assisted learning produces surgeons who operate better or patients who fare better. Finally, the perspective adopted is explicitly that of a surgical trainee, which is a strength for the stated purpose and a source of bias for any other, and training contexts differ substantially between health systems.

CONCLUSION

Artificial intelligence will not replace the surgical resident, and that framing has distracted from a more consequential question: whether the resident who trains alongside these tools will emerge as capable as the one who trained without them. The evidence suggests the answer is determined not by the technology but by the sequence in which it is used. AI tutoring added to expert instruction outperforms either alone, whereas AI tutoring substituted for expert instruction produces gains that do not survive blinded human assessment. Routine AI assistance in a procedural specialty measurably eroded unassisted performance in experienced operators. Model accuracy is highest exactly where the trainee needs it least and lowest exactly where the trainee needs it most. These are cognitive prostheses: they amplify a surgeon who already possesses judgement and conceal the absence of judgement in one who does not. The trainee's obligation is to build the underlying competence first and keep exercising it, to verify what matters at its source, to protect patient data, to disclose use honestly, and to insist that programmes teach these skills rather than leaving each resident to improvise them. The apprenticeship is not obsolete. It has acquired a very capable, very fluent and occasionally very wrong new participant, and the trainee's task is to remain the one who decides.

Funding: No funding sources

Conflict of interest: None declared

Ethical approval: Not required

REFERENCES

- George BC, Bohnen JD, Williams RG, Meyerson SL, Schuller MC, Clark MJ, et al. Readiness of US general surgery residents for independent practice. *Ann Surg.* 2017;266(4):582-94.
- Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84(2):273-8.
- Rubalcava NS, Guetter CR, Kapani N, Quiñones PM. How artificial intelligence is expected to transform surgical training. *Bull Am Coll Surg.* 2023;108(8).
- Kirubakaran A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. *J Surg Educ.* 2022;79(2):500-15.
- Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. *Res Integr Peer Rev.* 2019;4:5.
- Pedrett R, Mascagni P, Beldi G, Padoy N, Lavanchy JL. Technical skill assessment in minimally invasive surgery using artificial intelligence: a systematic review. *Surg Endosc.* 2023;37(10):7412-24.
- Kim JWB, Chen JT, Hansen P, Shi LX, Goldenberg A, Schmidgall S, et al. SRT-H: a hierarchical framework for autonomous surgery via language-conditioned imitation learning. *Sci Robot.* 2025;10(104):eadt5254.
- Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. *J Educ Eval Health Prof.* 2024;21:17.
- Tsai CY, Hsieh SJ, Huang HH, Deng JH, Huang YY, Cheng PY. Performance of ChatGPT on the Taiwan urology board examination: insights into current strengths and shortcomings. *World J Urol.* 2024;42(1):250.
- Kollitsch L, Eredics K, Marszalek M, Rauchenwald M, Brookman-May SD, Burger M, et al. How does artificial intelligence master urological board examinations? A comparative analysis of different large language models' accuracy and reliability in the 2022 in-service assessment of the European Board of Urology. *World J Urol.* 2024;42(1):20.
- Talyshinskii A, Juliebo-Jones P, Hameed BMZ, Naik N, Adhikari K, Zhanbyrbekuly U, et al. ChatGPT as a clinical decision maker for urolithiasis: compliance with the current European Association of Urology guidelines. *Eur Urol Open Sci.* 2024;69:51-62.
- Khalpey Z, Kumar U, King N, Abraham A, Khalpey AH. Large language models take on cardiothoracic surgery: a comparative analysis of the performance of four models on American Board of Thoracic Surgery exam questions in 2023. *Cureus.* 2024;16(7):e65083.
- Hetz MJ, Carl N, Haggenmüller S, Wies C, Kather JN, Michel MS, et al. Superhuman performance on urology board questions using an explainable language model enhanced with European

- Association of Urology guidelines. *ESMO Real World Data Digit Oncol.* 2024;6:100078.
14. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One.* 2020;15(2):e0229596.
15. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. *JAMA Netw Open.* 2022;5(2):e2149008.
16. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in surgical curriculum design and unintended outcomes for technical competencies in simulation training. *JAMA Netw Open.* 2023;6(9):e2334658.
17. Yilmaz R, Bakhaidar M, Alsayegh A, Abou Hamdan N, Fazlollahi AM, Tee T, et al. Real-time multifaceted artificial intelligence vs in-person instruction in teaching surgical technical skills: a randomized controlled trial. *Sci Rep.* 2024;14:15130.
18. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, Wang A, Makhani HH, Alsayegh A, et al. Effect of feedback modality on simulated surgical skills learning using automated educational systems: a four-arm randomized control trial. *J Surg Educ.* 2024;81(2):275-87.
19. Giglio B, Albeloushi A, Alhaj AK, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial intelligence-augmented human instruction and surgical simulation performance: a randomized clinical trial. *JAMA Surg.* 2025;160(9):993-1003.
20. Boal MW, Anastasiou D, Tesfai F, Ghamrawi W, Mazomenos E, Curtis N, et al. Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review. *Br J Surg.* 2024;111(1):znad331.
21. Winkler-Schwartz A, Yilmaz R, Mirchi N, Bissonnette V, Ledwos N, Siyar S, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw Open.* 2019;2(8):e198363.
22. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *NPJ Digit Med.* 2022;5:54.
23. Lavanchy JL, Zindel J, Kirtac K, Twick I, Hosgor E, Candinas D, et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Rep.* 2021;11:5197.
24. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): a surgical activity dataset for human motion modeling. In: *MICCAI Workshop M2CAI.* 2014;3:3.
25. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. Updated 2026. Available at: <https://www.icmje.org>. Accessed 20 August 2026.
26. Zielinski C, Winker MA, Aggarwal R, Ferris LE, Heinemann M, Lapeña JF Jr, et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Colomb Med (Cali).* 2023;54(3):e1015868.
27. Committee on Publication Ethics. Authorship and AI tools: COPE position statement. 2023. Available at: <https://publicationethics.org/cope-position-statements/ai-author>. Accessed on 20 August 2026.
28. Flanagan A, Kendall-Taylor J, Bibbins-Domingo K. Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. *JAMA.* 2023;330(8):702-3.
29. Cheng A, Calhoun A, Reedy G. Artificial intelligence-assisted academic writing: recommendations for ethical use. *Adv Simul (Lond).* 2025;10(1):22.
30. Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, et al. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *Lancet Gastroenterol Hepatol.* 2025;10(10):896-903.
31. Ray K. AI-assisted colonoscopy and risk of endoscopist deskilling. *Nat Rev Gastroenterol Hepatol.* 2025;22:672.
32. Abdunour RE, Gin B, Boscardin CK. Educational strategies for clinical supervision of artificial intelligence use. *N Engl J Med.* 2025;393(8):786-97.
33. Braunstein D, Farooq H, Paolino M, Moynihan A, Cahill RA. Comparative analysis of perceptions on artificial intelligence in surgery: a survey study among surgeons and medical students in Ireland. *Ir J Med Sci.* 2025;194(6):2341-9.
34. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. *Nat Med.* 2025;31:1873-81.
35. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: World Health Organization; 2024. Available at: <https://www.who.int/publications/i/item/9789240084759> Accessed on 20 August 2026.
36. Government of India. The Digital Personal Data Protection Act, 2023. New Delhi: Ministry of Electronics and Information Technology; 2023. Act No. 22 of 2023. Available at:

<https://www.indiacode.nic.in/indiacode/handle/123456789/22037> Accessed on 20 August 2026.

37. European Parliament, Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024 Jul 12;L 2024/1689. Available at: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> Accessed on 20 August 2026.
38. Indian Council of Medical Research. Ethical guidelines for application of Artificial Intelligence in Biomedical Research and Healthcare. New Delhi: Indian Council of Medical Research; 2023. Available at: https://www.icmr.gov.in/icmrobject/custom_data/pd

f/Ethical-guidelines/Ethical_Guidelines_AI_Healthcare_2023.pdf. Accessed on 20 August 2026.

39. Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. *Eur Urol.* 2024;85(2):146-53.

Cite this article as: Vinod A, Venkatesh V, John GM. Surgical training in the era of artificial intelligence and large language models: a narrative review from the perspective of a surgical resident. *Int Surg J* 2026;13:2376-85.