MARCINIAK, Kinga, SCIĄŻKO-GANCARCZYK, Sabina, WĘGRZYN, Jagoda, ŻUKOWIECKA-SĘGA, Dominika and GANCARCZYK, Maciej. From Virtual Tutors to Professional Identity: Generative AI and Large Language Models in Medical Education. *Journal of Education, Health and Sport*. 2025;86:66953. eISSN 2391-8306.
https://doi.org/10.12775/JEHS.2025.86.66953
https://apcz.umk.pl/JEHS/article/view/66953

The journal has had 40 points in Minister of Science and Higher Education of Poland parametric evaluation. Annex to the announcement of the Minister of Education and Science of 05.01.2024 No. 32318. Has a Journal's Unique Identifier: 201159. Scientific disciplines assigned: Physical culture sciences (Field of medical and health sciences); Health Sciences (Field of medical and health sciences).
Punkty Ministerialne 40 punktów. Załącznik do komunikatu Ministra Nauki i Szkolnictwa Wyższego z dnia 05.01.2024 Lp. 32318. Posiada Unikatowy Identyfikator Czasopisma: 201159. Przypisane dyscypliny naukowe: Nauki o kulturze fizycznej (Dziedzina nauk medycznych i nauk o zdrowiu); Nauki o zdrowiu (Dziedzina nauk medycznych i nauk o zdrowiu).
The authors declare that there is no conflict of interests regarding the publication of this paper.
Received: 28.11.2025. Revised: 08.12.2025. Accepted: 08.12.2025. Published: 13.12.2025.

From Virtual Tutors to Professional Identity: Generative AI and Large Language Models in Medical Education

Kinga Marciniak,
ORCID https://orcid.org/0009-0000-2322-9295
E-mail: kingamarciniak.kontakt@gmail.com
The HCP Hospital in Poznan, ul 28 Czerwca 1956 r. 194, Poznan, Poland

Jagoda Węgrzyn,
ORCID https://orcid.org/0009-0007-7426-8850
E-mail jagodaw613@gmail.com
Private Healthcare Facility Medical – M&S, Czeladź, Poland
Navigare Family Medicine Center, Dąbrowa Górnicza, Poland

Dominika Żukowiecka-Sęga, MD,
ORCID https://orcid.org/0009-0000-3426-7393
E-mail: dominikazukowiecka@gmail.com
Complex of Municipal Hospitals in Chorzów, ul. Strzelców Bytomskich 11, 41-500 Chorzów, Poland

Maciej Gancarczyk, MD,
ORCID https://orcid.org/0009-0004-3741-0254
Regional Specialist Hospital in Grudziądz, Poland

Sabina Ściążko-Gancarczyk, MD,
ORCID https://orcid.org/0009-0002-1738-3119
Regional Specialist Hospital in Grudziądz, Poland

Content

Abstract

Background:

Generative artificial intelligence (AI), particularly large language models (LLMs), is rapidly transforming medical education. These tools can act as virtual tutors, generate teaching materials and support clinical simulations, but they also raise concerns regarding accuracy, bias, academic integrity, data protection and professional identity formation.

Objective:

This narrative review aims to synthesise current knowledge on the use of generative AI - with a focus on LLMs - in medical education, identify key benefits and risks, and discuss implications for curricula and professionalism in medicine and other health professions.

Methods:

A literature search was conducted in PubMed/MEDLINE, Scopus and Web of Science in November 2025. Publications from 2021–2025 were included if they addressed applications of generative AI/LLMs in medical or health professions education, or discussed their impact on curricula, assessment, professionalism or future physician competencies. Original empirical studies, reviews, conceptual papers and policy/guideline documents were included. Data were extracted and synthesised narratively according to four domains: areas of application, benefits, risks/limitations and recommendations/frameworks.

Results:

Generative AI is being used to support self-directed learning (on-demand explanations, personalised practice questions), clinical simulations and virtual patients, faculty work in content generation and assessment, and research in medical education. Reported benefits include increased accessibility, personalisation, scalability and opportunities to practise clinical reasoning and communication in low-stakes environments. Major risks comprise hallucinations and bias, threats to academic integrity, potential deskilling and over-reliance, and privacy and confidentiality concerns. Emerging frameworks emphasise responsible use, including transparency of AI assistance, critical appraisal of outputs, proportional use aligned with learning goals, and strict data protection.

Conclusions:

Generative AI has substantial potential to enhance medical education, but its value depends critically on how it is implemented. Clear institutional policies, integration of AI literacy and professionalism into curricula, and sustained human oversight are essential to ensure that these tools complement rather than undermine core clinical competencies and the professional identity of future physicians.

Keywords:

generative artificial intelligence; large language models; ChatGPT; medical education; health professions education; curriculum development; clinical reasoning; academic integrity; professionalism; professional identity formation

Introduction

In recent years there has been a rapid development of artificial intelligence (AI) tools, and in particular of so-called generative AI based on large language models (LLMs). Models such as ChatGPT, Gemini and other conversational systems are capable of generating text, code, images and other content at a level that is often difficult to distinguish from human work. This phenomenon has important implications for many areas of healthcare systems - from diagnostics and research to medical education.

Medical education has long faced a number of challenges: the accelerating growth of knowledge, a limited number of clinical teachers, increasing student enrolment, and the need to develop not only theoretical knowledge, but also clinical reasoning, communication and professionalism. The COVID-19 pandemic further accelerated the digitalisation of education, popularising remote teaching, e-learning, high-fidelity simulations and hybrid models. In this context, generative AI has begun to be perceived as a potential “accelerator” of the transformation of medical education, offering, among others, personalised learning support, rapid generation of teaching materials, and simulations of patient - clinician dialogues.

Preliminary studies indicate that large language models achieve performance comparable to, or higher than, medical students on some test-based examinations and are able to generate examination questions, short clinical vignettes and exemplar diagnostic and therapeutic plans. From the student’s perspective, they may serve as a “virtual tutor”, helping to explain difficult concepts, generate quizzes, flashcards for revision, or case scenarios. For teachers, they constitute a potential tool to reduce workload, supporting the preparation of materials, variants of cases for discussion, and even the provision of preliminary feedback on written assignments.

At the same time, enthusiasm about generative AI is tempered by numerous concerns. These models may generate incorrect or “hallucinatory” responses, perpetuate existing biases embedded in training data, and their functioning largely resembles a “black box”. Serious doubts arise regarding academic integrity - from plagiarism and inappropriate assistance in writing coursework, to difficulties in assessing students’ genuine competences. Another important issue is the protection of sensitive data and the risk of entering clinical information into public models in a way that could allow patient re-identification.

An additional challenge is the impact of generative AI on the development of professional identity and professionalism among future physicians. Questions arise as to how the use of tools capable of rapidly summarising the literature, generating differential diagnoses or proposing management plans may modify learning processes, clinical reasoning and decision-making. On the one hand, generative AI may support the development of critical thinking if used reflectively and under supervision. On the other hand, uncontrolled, “unreflective” reliance on model suggestions may lead to the erosion of some key competences.

In response to the dynamic development of these technologies, the first institutions and organisations are developing guidelines for the responsible use of AI in medical education. These cover, among others, the scope of permissible use of models by students, the requirement for transparent disclosure of AI assistance, standards for data protection and the inclusion of AI literacy in curricula. Despite the growing number of publications, reports concerning generative AI remain fragmented, and the overall picture is ambiguous - ranging from highly enthusiastic narratives to strongly sceptical positions.

In this situation, there is a need for a synthetic literature review that systematically presents current and potential applications of generative AI in medical education, identifies key benefits and risks, and discusses major implications for curricula in medicine and other health professions.

The aim of this article is to review the current state of knowledge on the applications of generative artificial intelligence - in particular large language models - in medical education. The article focuses on (1) the main areas of use of generative AI in educational processes, (2) the potential benefits and limitations of these tools from the perspective of students and teachers, and (3) ethical, legal and professionalism-related challenges. On this basis, preliminary recommendations will be proposed for curriculum designers and medical education practitioners concerning the responsible implementation of generative AI into educational processes.

1. Methods

Type of article

This article is a narrative review. Its aim was to provide a synthetic overview of the current state of knowledge on the applications of generative artificial intelligence, in particular large language models (LLMs), in medical education, and to identify the main benefits, limitations and implications for medical curricula.

Literature search strategy

The review was based on a literature search in the following databases:

- PubMed/MEDLINE
- Scopus
- Web of Science

The search was conducted in November 2025, without geographical restrictions. Given the rapid pace of development of generative AI technologies, the search was limited to publications from 2021-2025, with the exception of key review articles and conceptual papers describing the fundamentals of LLMs.

An example search strategy for PubMed was as follows:

> ("generative artificial intelligence" OR "generative AI" OR "large language model*" OR "large-language model*" OR "LLM" OR "ChatGPT" OR "GPT-4" OR "foundation model*")
>
> AND
>
> ("medical education" OR "health professions education" OR "undergraduate medical education" OR "graduate medical education" OR "clinical training")

In Scopus and Web of Science, equivalent strategies were used, with adaptation of keywords and Boolean operators to the specifics of each database. In addition, a “snowballing” approach was applied by analysing reference lists in identified review articles and key conceptual works in order to find further relevant publications.

Inclusion and exclusion criteria

The review included publications that met the following criteria:

Type of publication:

- original research articles (quantitative, qualitative, mixed-methods),
- review articles,
- conceptual/viewpoint/commentary articles,
- guidelines and policy/position statements relevant to medical education.

Topical scope:

- application of generative AI, in particular large language models (e.g., ChatGPT, other LLMs), in medical education or health professions education (medical students, students in related disciplines, residents, physicians in postgraduate training),

or

- discussion of the impact of generative AI on curricula, assessment, professionalism and the competences of future physicians.

Language of publication:

- English (optionally: and Polish).

The following were excluded from the review:

- publications not directly related to medical education (e.g., purely clinical applications of AI without an educational component),
- conference reports without full text, abstracts without a full article, press releases, blog posts,
- studies focused exclusively on classical AI/ML methods, without reference to generative AI or LLMs.

Study selection procedure

The selection process followed two stages:

1. Title and abstract screening - after removal of duplicates, titles and abstracts of the identified records were screened for compliance with the inclusion criteria. Publications clearly not meeting the criteria were excluded at this stage.
2. Full-text assessment - for studies retained after initial screening, full texts were retrieved and their content was assessed against the detailed inclusion and exclusion criteria. In cases of doubt regarding eligibility, an inclusive approach was adopted: an article was included if it contributed important information to at least one of the analysed domains (applications, benefits, risks, implications for education).

No formal, quantitative assessment of the methodological quality (risk of bias) of individual studies was conducted, which is in line with the narrative review design. However, when interpreting the findings, the type of study, its scope and methodological limitations reported by the authors were taken into account.

Data extraction and synthesis
From the included publications, the following information was extracted:

- bibliographic data (author, year, journal, country/region),
- type of article (original study, review, commentary, guideline),
- level of education (undergraduate, postgraduate, other health professions),
- nature of the application of generative AI/LLMs (e.g., support for self-directed learning, clinical simulations, generation of teaching materials, assessment),
- key findings, including:
    - reported benefits and possible positive effects,
    - identified risks, limitations and challenges,
    - authors' recommendations for practical implementation of AI,
      main conclusions and identified research gaps.

Given the heterogeneous nature of the included works (quantitative and qualitative studies, reviews, conceptual papers), a narrative synthesis was adopted. The findings were organised thematically around:

- areas of application of generative AI in medical education,
  benefits and potential positive effects,
  risks, limitations and challenges (including ethical, legal and professionalism-related issues),
- recommendations and proposed frameworks for implementing generative AI into curricula.

On this basis, synthetic conclusions were formulated and potential directions for future research were identified.

Limitations of the review
This review is narrative in nature and does not meet the formal criteria of a systematic review. Despite the use of an extensive search strategy, there remains a risk of omitting some relevant publications, particularly those published outside major databases or in languages other than English. Furthermore, the rapid development of generative AI means that the state of knowledge presented in this article reflects the situation at the time of the search (date) and may require updating in the near future.

2. Definitions and classification of generative artificial intelligence in medical education

Artificial intelligence (AI) can be broadly defined as a set of computational methods that enable the performance of tasks that typically require human cognitive abilities - such as learning from data, pattern recognition, reasoning and decision-making. In the literature on medical education, two broad classes of systems are traditionally distinguished:

- "classical" systems based on rules or predictive models (machine learning),
- newer generative systems capable of creating novel content that was not explicitly present in the training data (Hersh, W., 2025).

Classical (predictive) AI systems

In the "classical" sense, AI in medicine includes, among others: rule-based expert systems, machine-learning models (e.g., logistic regression, decision trees, random forests, neural networks), as well as deep classification models, for example in radiology or dermatology. Their main function is prediction:

- assigning a patient to a given class (e.g., diseased/healthy),
- estimating risk (e.g., probability of complications),
- recognising patterns in images or biological signals.

In medical education, such tools have thus far been used mainly for:

- adaptive testing systems and intelligent tutoring (adjusting the difficulty level of tasks),
- analysis of examination results and identification of areas requiring support,
- diagnostic simulations based on predefined decision algorithms (Thompson, R. et al.; 2025).

Generative artificial intelligence

Generative artificial intelligence (generative AI) refers to models that learn the distribution of input data (e.g., text, images, audio) and can subsequently generate new examples that are statistically consistent with that distribution. Hersh proposes a distinction between:

- predictive AI -which, on the basis of input data, predicts a label, a numerical value or a probability (e.g., risk of death, diagnosis based on an image), and
- generative AI - which creates new content (text, images, code, multimodal responses) used, for example, to summarise information, answer questions or simulate interactions (Hersh, W., 2025).

In medical education, generative AI is currently associated primarily with large language models (LLMs) and related models that generate images and multimodal content.

Large language models (LLMs)

Large language models are deep neural networks trained on very large text corpora that can predict the next token in a sequence and, consequently, generate coherent natural language text. Examples include GPT-4 and more recent models from the GPT family, Claude, Gemini and a range of open-source models.

In the context of medical education, LLMs can:

- answer examination and clinical questions at a level comparable to candidates taking specialty examinations,
- generate summaries of scientific articles and textbooks,
- create clinical cases, OSCE scenarios and test questions,
- simulate conversations with a patient or "virtual patient" in educational settings,
  act as a "virtual tutor" that explains difficult concepts in real time and asks follow-up questions.

Thompson et al. emphasise that LLMs represent a qualitatively new category of tools in medical education - in contrast to earlier, narrower AI applications, they can support many elements of the educational process simultaneously, from self-directed learning to the creation of teaching materials (Thompson, R. et al.; 2025).

Generative image models and multimodal models

Alongside LLMs, generative image models (e.g., based on generative adversarial networks or diffusion models) and multimodal models that integrate text with images, clinical data or signals are being intensively developed. In medicine, they are used, among others, to:

- synthetically generate radiological or dermatological images,
- augment training datasets,
- illustrate clinical cases,
- create interfaces that simultaneously "understand" text and images (e.g., an imaging report together with the corresponding image).

In medical education, potential applications include:

- generating diverse examples of imaging cases for learning diagnostic skills,
- building interactive atlases and simulators (e.g., anatomy, radiology, dermatology),
- practising the description of imaging studies and clinico-radiological correlation (Hersh, W., 2025).

3. Areas of application of generative AI in medical education

Support for individual student learning

Early empirical studies and reviews suggest that large language models (LLMs) are most commonly used as tools for self-directed learning and "just-in-time" support at the point of study. A recent scoping review of LLMs in medical education identified three dominant patterns of use: (1) clarification and elaboration of complex concepts, (2) generation of learning materials (questions, cases, summaries) and (3) interactive practice environments (Luordo, D. et al., 2025).

From the student perspective, LLMs such as ChatGPT are frequently described as "on-demand explainers", capable of rephrasing difficult topics in simpler language, offering stepwise explanations or comparing alternative frameworks (e.g. differential diagnoses, management options). These uses in undergraduate medical education, emphasising that ChatGPT can scaffold learning in both preclinical and clinical courses, provided that students are encouraged to critically appraise answers against trusted sources (Almansour, M.; Alfhaid F.; 2024). Studies evaluating LLM performance on high-stakes examinations (e.g. USMLE) show that current models can reach or exceed passing thresholds, underscoring their potential as content-knowledge resources but also raising concerns about overreliance and academic integrity (Scherr R, et al., 2023).

Generative AI is also being used to create personalised practice material. Students commonly use LLMs to generate multiple-choice questions (MCQs), short-answer prompts, and brief clinical vignettes aligned with specific learning objectives or examination blueprints (Luordo, D. et al., 2025). Dedicated tools such as Anki Tagger leverage LLMs to map large pre-existing flashcard libraries to local curricula, allowing learners to prioritise cards that match their school's objectives and thereby structure spaced-repetition learning more efficiently (Pendergrast T, Chalmers Z., 2023). Case reports and small studies suggest that such AI-assisted curation may reduce the time needed to search for relevant resources and improve perceived alignment between self-study and course assessments, although robust outcome data remain limited (Luordo, D. et al., 2025).

Beyond question generation, several authors describe students using ChatGPT and similar systems as “virtual study partners” to simulate oral viva questions, construct comparative tables (e.g. between diseases or drugs), generate mnemonics and produce tailored study plans based on learner-specified timelines and goals (Skryd A, Lawrence K., 2024). However, existing reviews consistently emphasise the risks of factual errors, superficial explanations and hallucinated citations; they recommend that LLM outputs be treated as prompts for further reading rather than authoritative sources, and that students receive explicit training in checking references and cross-validating information (Metze K., et al., 2023).
Overall, current evidence suggests that generative AI can enhance individual learning by providing flexible, low-threshold access to personalised explanations and practice materials. At the same time, without explicit guidance and assessment policies, there is a risk that students may substitute AI-generated answers for active engagement with primary literature and neglect deeper conceptual understanding (Metze K., et al., 2024).

Clinical simulations and virtual patients

A second rapidly evolving area is the use of generative AI to power clinical simulations and virtual patients. Scherr and colleagues demonstrated that ChatGPT-3.5 can be configured to run interactive simulations for early clinical education, including advanced cardiac life support and intensive-care scenarios. The model was able to respond dynamically to student input, adapt the course of the case and provide post-scenario feedback, thereby offering a level of interactivity beyond static vignette-based questions (Scherr R., 2023). The authors highlighted advantages such as scalability and low marginal cost per additional scenario, but also noted concerns around medical accuracy, reproducibility of outputs and the need for careful prompt engineering.
More structured implementations of LLM-based virtual patients are now emerging. Sheth et al. evaluated an LLM-driven “OSCE-style” virtual patient for history-taking practice in undergraduate medical education. Following participation in simulated interviews with the AI patient, students reported a 14.6% increase in self-rated comfort with history-taking. Qualitative data indicated that learners valued the flexibility, detailed (if sometimes repetitive) feedback and the opportunity to practise in a “judgement-free” environment. However, the system was perceived as lower fidelity than encounters with standardised patients, and some feedback was deemed less clinically relevant than that provided by human preceptors (Sheth, U. et al., 2025).
Taken together, available data suggest that generative-AI-driven virtual patients are feasible and acceptable as adjuncts to traditional simulation modalities, particularly for low-stakes practice in history-taking and basic clinical reasoning. They are not yet ready to replace high-fidelity simulations or standardised patients in summative assessment, and their use requires clear guidance on case vetting, faculty oversight and safeguarding of learner and patient data (Laupichler MC. et al., 2023).

Teacher support: content generation and assessment

Generative AI is increasingly explored as a tool to reduce faculty workload in the design of teaching materials and assessment. A systematic review of LLM-generated medical examination questions identified studies using ChatGPT-3.5 and GPT-4 to create MCQs; expert raters generally judged these items as broadly comparable in quality to human-written questions, and in some cases they were successfully used in formative assessments. Nevertheless, a considerable proportion of questions required revision to address factual errors, implausible distractors or inappropriate difficulty. In a comparative study, Laupichler et al. showed that although students performed similarly on ChatGPT-generated and faculty-written neurophysiology MCQs, the discrimination indices of AI-generated items were significantly lower, suggesting reduced ability to differentiate between stronger and weaker candidates. Beyond MCQs, narrative reports describe educators using generative AI to draft case-based learning scenarios, OSCE station scripts and variants of clinical vignettes based on human-written "seed" cases. These applications are regarded as promising for expanding the pool of available cases and supporting curricular renewal, but they also raise concerns about the introduction of subtle errors, potential leakage of test content and drift from the intended blueprint if outputs are not systematically reviewed (Komasawa N, Yokohira M., 2025).

Generative AI is also being applied to assessment and feedback. Studies on AI-assisted scoring of OSCE performance have shown good agreement with human examiners and gains in inter-rater reliability, as well as substantial time savings when AI is used to support grading. However, discrepancies persist in domains such as communication and professionalism, and authors consistently emphasise that AI should function as a supplementary evaluator rather than a replacement for human assessors. Similarly, LLMs are being piloted to provide formative feedback on short written assignments (e.g. clinical summaries, reflective pieces), with students reporting such feedback as timely and helpful for improving structure, clarity and completeness. At the same time, reviews highlight important risks, including reinforcement of bias, misinterpretation of context and overconfident but erroneous evaluations. Ethical guidance therefore recommends that AI-mediated feedback be clearly labelled, that learners be explicitly trained to appraise it critically, and that high-stakes judgements about competence remain the responsibility of qualified faculty (Hersh, W., 2025).

Support for research in medical education

In addition to teaching and assessment, generative AI is increasingly used to support research in medical education. Narrative and scoping reviews describe several potential roles: assisting with initial literature scoping, suggesting search terms, summarising article abstracts, helping to structure manuscripts and generating preliminary versions of survey items or interview guides (Komasawa N, Yokohira M., 2025). These uses are reported to save time and lower the threshold for novice researchers, particularly non-native English speakers.

However, a growing body of evidence warns against relying on LLMs as bibliographic tools. Metze demonstrated that ChatGPT frequently fabricated references or combined details from multiple real articles, and that the error rate increased with topic complexity (Metze K. et al., 2024). In a subsequent study of ChatGPT-4, the same group showed that bibliographic outputs were unstable across sessions and versions, with persistent hallucinations and only sporadic "alerts" indicating uncertainty (Metze K., 2024).

Similar concerns are echoed in reports from other clinical domains, which conclude that LLMs are not suitable as primary tools for systematic literature searches or evidence synthesis.
With respect to instrument development, authors have begun exploring LLMs for generating candidate items for questionnaires, checklists or rating scales. Insights from studies on AI-generated exam questions suggest that while models can rapidly produce large numbers of plausible-looking items, these require thorough expert review, pilot testing and psychometric validation. There are also unresolved questions about intellectual property and authorship when AI contributes substantively to item wording or manuscript drafting (Komasawa N, Yokohira M., 2025).
Finally, many commentaries emphasise the implications of generative AI for research integrity. Potential risks include plagiarism (through over-reliance on AI-written text), inadvertent inclusion of fabricated data or references, and opacity about the extent of AI assistance. Consequently, recent guidelines and narrative reviews call for transparency in reporting AI use, explicit institutional policies, and integration of "AI literacy" (including understanding hallucinations, bias and limitations) into the training of clinician-educators and health professions education researchers (Komasawa N, Yokohira M., 2025).

4. Benefits and potential of generative AI in medical education

The spread of large language models (LLMs), such as ChatGPT and others, has significantly changed the landscape of medical education. Literature reviews indicate that generative AI is already being used in the training of medical students and physicians, among others as a virtual tutor, a tool for personalized learning, a generator of educational materials, and a supplement to clinical simulations (Wong RSY., 2024).
At the same time, authors emphasize that evidence for long-term effectiveness is still limited, and the potential of this technology can only be realized under conditions of critical supervision by educators and within clear ethical frameworks (Cione, Nicholas J., 2025).

Personalization and accessibility
One of the most frequently highlighted advantages of generative AI is its ability to act as an "on-demand tutor" – an interactive teaching assistant available whenever needed. In a scoping review on personalized learning using LLMs in medical education, a range of applications was described, including:

- intelligent tutoring (answering questions, simplifying complex concepts),
- generating personalized learning plans,
- adaptive tests with immediate feedback,
- and summarizing texts with the level of difficulty adjusted to the student's knowledge.

These studies show that LLMs can reduce information overload for students, shorten the time needed to prepare for classes, and enable learning at one's own pace (Vrdoljak J., 2025).
In case reports and pilot studies, generative "teaching assistants" embedded as chat tools in e-learning environments have been described. These models answer students' questions, explain doubts step by step, and propose additional tasks tailored to the level of proficiency - for example, more difficult questions for advanced learners and more detailed hints for those who are struggling (Hui, Z., et al., 2025).

Preliminary results indicate improved subjective feelings of support, better theoretical test scores in groups using ChatGPT in supported problem-based learning (PBL), and a positive evaluation of the AI tutor's "responsiveness" (Thesen, T., Park, S.H., 2025).
Accessibility and inclusivity are also important aspects. Generative AI can adapt the way content is presented to a preferred learning style (text explanations, diagrams, bullet points, test questions), and it can simplify or translate text into other languages, which may be particularly helpful for students studying in a language that is not their mother tongue. Although empirical research on reducing language barriers is still limited, numerous conceptual and review papers point to the potential of LLMs to support culturally and linguistically diverse students and to promote educational equity, provided that appropriate infrastructure and digital capabilities are ensured (Simoni, J. et al., 2025).
Crucially, generative AI responds particularly well to the challenges associated with large student cohorts. Using LLMs as a first line of educational support can relieve faculty of some repetitive tasks (answering similar questions, explaining basic concepts), allowing teachers to focus on more complex activities: working with difficult cases, shaping professional attitudes, or providing individual mentoring (Vrdoljak J., 2025).

Strengthening clinical reasoning (when used appropriately)
Clinical reasoning - encompassing data gathering, formulating and verifying diagnostic hypotheses, and planning therapy – is one of the key, yet most difficult to implement, components of physician training. A number of studies indicate that generative AI can support the development of these competencies if it is appropriately integrated into the curriculum (Vrdoljak J., 2025).
First, LLMs can serve as tools for training analytical thinking and generating differential diagnoses. Qualitative studies have described the use of ChatGPT to generate and develop clinical cases for students - the model suggested possible diagnoses, subsequent diagnostic steps, and comments on the reasoning process, which were then verified and refined by teachers (Wong K., et al., 2024). In another study, using ChatGPT "on ward rounds" showed that the tool could help the student team with differential diagnosis, identifying knowledge gaps, and generating additional questions for bedside discussion (Skryd A, Lawrence K., 2024).
In quantitative studies on the usefulness of ChatGPT as a "diagnostic companion", it has been shown that the model is usually reliable in ruling out some diagnoses and suggesting logical next diagnostic steps, and its answers are rated as clear and well suited for educational purposes. The authors emphasize that such a system may be valuable as a tool for differentiating disease causes and structuring clinical thinking - but not as an independent "diagnostician" (Hadi A. et al., 2024).
Second, generative AI is increasingly being used to create simulated patient encounters and virtual patients. Studies on building "virtual patients" using LLMs show that models can generate coherent, varied clinical scenarios, respond to students' questions in a way that resembles natural conversation, and modify the course of the case depending on the learner's decisions (Öncü, S., et al., 2025).
Pilot studies indicate that such simulations increase student engagement and allow them to practice history-taking, elements of clinical reasoning, and communication with patients in safe and repeatable conditions (Öncü, S., et al., 2025).

In work on AI-powered standardized patients (based on ChatGPT-4o), it has been shown that generative AI can realistically portray patients in different scenarios, provide students with individualized feedback, and enable a practically unlimited number of repetitions, which is difficult to achieve with actors or traditional simulators (Öncü, S., et al., 2025).
Articles describing the use of generative AI to design clinical cases and OSCE scenarios also emphasize that this tool makes it possible to quickly generate many variants of the same problem at different levels of difficulty, which supports the grading of challenges and better alignment with students' levels (Wong K., et al., 2024).
At the same time, numerous publications warn against the risks of uncritical reliance on model suggestions. They point to hallucinations (confident but incorrect answers), training data bias, the "automation" effect (a tendency to accept AI proposals without sufficient scrutiny), and the potential weakening of independent reasoning if the tool is used mainly to "provide ready-made answers" (Hadi A. et al., 2024).
Therefore, the role of generative AI in developing clinical reasoning should be defined as a tool for practice, provoking discussion, and generating counter-arguments, rather than as an authoritative source of truth. This requires active moderation by educators, incorporating elements of critical appraisal of AI responses, and clearly communicating the limitations of these systems to students.

5. Risks, limitations and the "dark sides"

Generative AI in medical education carries enormous potential, but at the same time comes with a range of risks: from hallucinations and bias, through problems with academic integrity, to erosion of competencies and threats to the confidentiality of patient data. The literature increasingly emphasizes that the way AI is implemented is just as important as the technology itself – without clear frameworks and oversight, models may weaken rather than strengthen the quality of education and patient safety.

Hallucinations, factual errors and bias

In the literature, it has quickly become accepted that hallucinations and factual errors are not a side effect, but a structural limitation of large language models. Yang and colleagues, in a review for the Annual Review of Biomedical Data Science, describe numerous examples in which LLMs generate plausible but partly incorrect clinical answers, false summaries of study results, and completely fabricated citations (including PMID numbers). Such "confabulations" appear even when the model correctly answers USMLE-type test questions, which further highlights the gap between the correctness of answers and the quality of reasoning (Yang Y et al., 2025).
Similar conclusions emerge from the review by Usuyama and colleagues, who summarize applications of LLMs in biomedical NLP: high performance on benchmarks coexists with local errors in the interpretation of medical documentation, incorrect reasoning in the presence of negations, and inconsistent use of clinical terminology (Usuyama N, et al., 2025).
In the educational context, this means that a student working with genAI may receive partially incorrect but stylistically "perfect" explanations that are difficult to challenge without guidance.
Issues of bias are just as serious.

Yang and colleagues describe how LLMs may assign diseases more frequently to specific racial or ethnic groups, reproducing patterns present in the training data (Yang Y et al., 2025).
An empirical example is provided by the work of Chan and Kwek, who analysed how different models assign cardiovascular risk to patients with identical clinical profiles but different sociodemographic characteristics - demonstrating substantial discrepancies and a potential to reproduce existing health inequities. In the literature on the use of LLMs in healthcare, the most frequently cited ethical risks are: lack of transparency, reinforcement of bias, hallucinations, and difficulties in assigning responsibility for errors (Chan JTN, Kwek RK., 2025). The authors stress that these problems are particularly sensitive in medical education: students are still developing their "epistemic compass" and may place excessive trust in models that speak in a "confident tone".

Academic integrity and plagiarism

With the advent of generative AI, the boundary between acceptable tool-supported work and impermissible assistance has become hard to draw clearly. Jaleel and colleagues, in an article in Frontiers in Education, show that among medical students ChatGPT is perceived both as a "virtual tutor" and as an "essay-writing machine", which generates numerous dilemmas around authorship, transparency, and the assessment of learning outcomes (Jaleel A, et al.).
Ekaterina and colleagues, in the paper Academic Integrity Within the Medical Curriculum in the Age of Generative Artificial Intelligence, argue that traditional definitions of plagiarism – focused mainly on copying other authors' text - are insufficient when a student generates an assignment using a model that is not an "author" in the classical sense of the term (Ekaterina K. 2025). The authors propose treating such use of genAI to produce assessed work as a new form of academic integrity violation.
At the systemic level, an important point of reference has been the position of the Australian regulator TEQSA. In its document "Gen AI - academic integrity and assessment reform", TEQSA stresses that there are no reliable technical methods for detecting genAI use in student texts and recommends shifting the emphasis away from "catching cheaters" towards redesigning assessment - in the direction of tasks based on oral work, demonstrations of skills, and authentic clinical contexts (TEQSA, 2025). These warnings have been widely covered in the media; The Australian reported, among other things, TEQSA's view that "cheating with AI in online assignments is practically impossible to detect", and described universities' responses (for example, the "dual-track" assessment model at the University of Sydney: free use of AI in take-home assignments alongside a ban in invigilated exams).
In parallel, medical schools are introducing their own policies. SUNY Downstate requires students to declare AI use (name of the tool, URL, the prompts and generated responses), and the absence of such information is treated as a breach of academic integrity rules. The Jacobs School of Medicine in Buffalo has adopted a formal "Generative Artificial Intelligence Use Policy for Medical Students", which regulates acceptable forms of AI use in pre-clinical and clinical courses. LSU Health New Orleans likewise lists a "Policy on Use of Generative AI by Medical Students" in its catalogue of binding regulations for the MD program, illustrating that the topic has become an integral component of institutional governance in medical education.

Erosion of competencies (“deskilling”) and over-reliance

Another concern relates to the erosion of clinical competencies as a result of long-term reliance on AI systems. Natali and colleagues, in a mixed-methods analysis devoted to “AI-induced deskilling in medicine”, show that the literature already contains numerous warning signals: from reduced diagnostic vigilance, through a decline in manual practice, to the gradual loss of the habit of independently searching for information (Natali C, et al., 2025).

A strong empirical example comes from the study by Budzyń and colleagues in The Lancet Gastroenterology & Hepatology, where experienced endoscopists, after several months of working with an AI system supporting colonoscopy, were worse at detecting precancerous lesions when the AI was temporarily turned off (“the Google Maps effect” for clinical vision) (Budzyń K, et al., 2025).

In medical education, this problem concerns above all critical thinking and independent clinical reasoning. Wong, in the commentary “ChatGPT in Medical Education: Promoting Learning or Killing Critical Thinking?”, points out that delegating analyses to ChatGPT too often may limit students’ creativity and their tendency to engage in deeper reflection on a clinical case (Wong RSY., 2024).

Privacy, data security and patient confidentiality

From the perspective of medical education, one of the most important limitations is the absolute prohibition on entering identifiable patient data into public models. Zhui and colleagues, in the article Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education, propose, among other things, the principle of “data minimization” and a clear distinction between tools run within institutional infrastructure and commercial cloud-based models over which the institution has no control (Zhui L, et al., 2024).

At the system level, similar emphases appear in the AI Code of Conduct for Health and Medicine developed by the National Academy of Medicine, where privacy, data protection and auditability of information flows are indicated as preconditions for the responsible implementation of AI in healthcare and clinical education.

In medical schools’ policies, the wording is becoming increasingly precise. The Jacobs School of Medicine in Buffalo, in its document on the use of Microsoft Copilot, explicitly states that category 1 data (protected PII, regulated data) may never be entered into any generative tools; only internal data (after authentication within the university domain) and public data are permissible.

In teaching practice, this translates into several simple but crucial rules: students should work with synthetic or heavily de-identified cases, use models deployed in a secure university environment, and clearly understand that any entry of identifiable patient data into a public AI chat constitutes a potential breach of confidentiality - with both legal and ethical consequences.

6. Professionalism, ethics and physician identity in the era of generative AI

The spread of generative artificial intelligence is changing not only how students, physicians and other health professionals learn, but also how we understand what professionalism and physician responsibility mean. Language models today can pass high-stakes exams, write technically correct scientific texts and propose diagnostic and therapeutic plans. This does not mean, however, that “knowledge” and “competence” are becoming obsolete - quite the opposite (Mbakwe AB et al., 2023).

The knowledge of medical students increasingly resembles a competence in working with information rather than simple memory: it includes formulating appropriate questions for the system, recognizing typical model errors (hallucinations, gaps, biases), verifying answers in reliable sources, and integrating AI suggestions with one’s own clinical reasoning. From the perspective of professionalism, this shifts the emphasis from the question “what does the student know?” to “how does the student use the available tools when a clinical decision has to be made?”. In this new view, an “AI-ready physician” is not someone who knows all the answers, but someone who can collaborate responsibly with technology while remaining ultimately accountable for the decision.

How to teach responsible use of AI

Since most students are already using GenAI tools, the key question is not so much “whether” but “how” to teach them to do so. Education should go beyond technical instructions and become part of ethical and professional formation. Universities and organizations such as the AAMC propose principles for the responsible use of AI, emphasizing, among other things, the need for human-centered care, privacy protection and transparency. In practice, this can be implemented through: introductory courses on AI in medicine, exercises in critical analysis of AI outputs, project assignments that require describing how AI was used, and clinical scenarios in which AI acts as an “additional team member” and students must justify when and why they disagree with its recommendations.

A useful teaching tool can be a simple model of four principles for responsible AI use by medical students: (1) transparency - always disclose that you are using AI and in what way, (2) critical thinking - treat model outputs as hypotheses, (3) proportionality - adjust the level of AI use to the purpose of the task and your stage of training, and (4) data protection - do not enter patient data or other sensitive data into public models.

The role of disclosure of AI use in assignments and examinations

Disclosure of AI use is one of the central ethical issues in medical education. Increasingly, universities are moving away from a simple ban towards an approach of “you may use AI, but you must disclose it and meet specific conditions”. A lack of disclosure is then treated similarly to plagiarism or other forms of misconduct.

The role of transparency is multidimensional: in the educational sphere, it allows teachers to realistically assess the student’s own contribution and to better design assignments; in the ethical sphere, it builds a culture of responsibility and trust; in the clinical sphere, it trains the habit of clearly explaining to the patient what role the algorithm played in the diagnostic and therapeutic process.

In practice, disclosure can take simple forms: a brief note in a written assignment, an "AI use statement" section, or discussing in class which parts of the work were produced with AI support. Such practices not only reduce the risk of misuse, but also promote responsible technology use as a core element of the future physician's professional identity.

7. Conclusions

Generative artificial intelligence, and large language models in particular, are already reshaping many aspects of medical education - from self-directed learning and clinical simulations to assessment and educational research. The available literature suggests that, when thoughtfully integrated, these tools can enhance personalisation, increase access to practice opportunities and support the development of clinical reasoning and communication skills. At the same time, they introduce a qualitatively new layer of complexity in domains such as academic integrity, data protection and the formation of professional identity.

The reviewed evidence highlights that the main risks are not purely technical, but educational and ethical: hallucinations and bias, potential erosion of independent reasoning, blurred boundaries between acceptable support and misconduct, and threats to patient confidentiality. These challenges call for a shift from ad hoc experimentation to deliberate, policy-driven implementation. Curriculum designers and educators need to define when and how AI may be used, how its role should be disclosed, and how students can be trained to critically appraise AI-generated content.

Looking ahead, responsible integration of generative AI into medical education will require continuous collaboration between educators, clinicians, learners, ethicists, legal experts and technology providers. Longitudinal, outcome-focused research is needed to evaluate the real impact of AI on learning, clinical performance and professionalism. If coupled with robust safeguards and explicit attention to professional values, generative AI can become a powerful catalyst for modernising medical education rather than a source of erosion of core competencies.

References

1. William Hersh. Generative Artificial Intelligence: Implications for Biomedical and Health Professions Education. Annual Review of Biomedical Data Science Volume 8, 2025, https://doi.org/10.1146/annurev-biodatasci-103123-094756 April 09, 2025
2. Thompson RAM, Shah YB, Aguirre F, Stewart C, Lallas CD, Shah MS. Artificial Intelligence Use in Medical Education: Best Practices and Future Directions. Curr Urol Rep. 2025 May 29;26(1):45. doi:10.1007/s11934-025-01277-1. PMID: 40439780; PMCID: PMC12122599.
3. Luordo, D.; Torres Arrese, M.; Tristán Calvo, C.; Shani Shani, K.D.; Rodríguez Cruz, L.M.; García Sánchez, F.J.; Lagares Gómez-Abascal, A.; Rubio García, R.; Delgado Jiménez, J.; Pérez Carreras, M.; et al. Application of Artificial Intelligence as an Aid for the Correction of the Objective Structured Clinical Examination (OSCE). Appl. Sci. 2025, 15, 1153. https://doi.org/10.3390/app15031153
4. Almansour, M.; Alfhaid F.; Generative artificial intelligence and the personalization of health professional education: A narrative review. PMID: 39093806 doi:10.1097/MD.0000000000038955; Aug 02, 2024;103 (31): e38955.
5. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT Interactive Medical Simulations for Early Clinical Education: Case Study. JMIR Med Educ. 2023 Nov 10;9:e49877. doi: 10.2196/49877. PMID: 37948112; PMCID: PMC10674152.
6. Pendergrast T, Chalmers Z. Anki Tagger: A Generative AI Tool for Aligning Third-Party Resources to Preclinical Curriculum. JMIR Med Educ. 2023 Sep 20;9:e48780. doi: 10.2196/48780. PMID: 37728965; PMCID: PMC10551781.
7. Skryd A, Lawrence K, ChatGPT as a Tool for Medical Education and Clinical Decision-Making on the Wards: Case Study, JMIR Form Res 2024;8:e51346, doi: 10.2196/51346, PMID: 38717811, PMCID: 11112466
8. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination. J Pediatr Surg. 2024 Jan;59(1):158. doi: 10.1016/j.jpedsurg.2023.08.018. Epub 2023 Aug 30. PMID: 37735041.
9. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic research with large language model ChatGPT-4: instability, hallucinations and sometimes alerts. Clinics (Sao Paulo). 2024 Oct 15;79:100409. doi: 10.1016/j.clinsp.2024.100409. PMID: 39413495; PMCID: PMC11530843.
10. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT Interactive Medical Simulations for Early Clinical Education: Case Study. JMIR Med Educ. 2023 Nov 10;9:e49877. doi: 10.2196/49877. PMID: 37948112; PMCID: PMC10674152.
11. Sheth, U.; Lo, M.; McCarthy, J.; Baath, N.; Last, N.; Guo, E.; Monteiro, S.; Sibbald, M. Understanding the Role of Large Language Model Virtual Patients in Developing Communication and Clinical Skills in Undergraduate Medical Education. Int. Med. Educ. 2025, 4, 39. https://doi.org/10.3390/ime4040039
12. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large Language Models in Medical Education: Comparing ChatGPT- to Human-Generated Exam Questions. Acad Med. 2024 May 1;99(5):508-512. doi: 10.1097/ACM.0000000000005626. Epub 2023 Dec 28. PMID: 38166323.

13. Komasawa N, Yokohira M. Generative Artificial Intelligence (AI) in Medical Education: A Narrative Review of the Challenges and Possibilities for Future Professionalism. Cureus. 2025 Jun 18;17(6):e86316. doi: 10.7759/cureus.86316. PMID: 40688936; PMCID: PMC12276793.
14. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic Research with ChatGPT may be Misleading: The Problem of Hallucination. J Pediatr Surg. 2024 Jan;59(1):158. doi: 10.1016/j.jpedsurg.2023.08.018. Epub 2023 Aug 30. PMID: 37735041.
15. Wong RSY. ChatGPT in medical education: promoting learning or killing critical thinking? Education in Medicine Journal. 2024;16(2):177–183. https://doi.org/10.21315/eimj2024.16.2.13
16. Cione, Nicholas J.; Pillow, Malford T. MD, MEd. Creating More Reliable Large Language Models in Medical Education. Academic Medicine 100(9):p 1001, September 2025. | DOI: 10.1097/ACM.0000000000006058
17. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare (Basel). 2025 Mar 10;13(6):603. doi: 10.3390/healthcare13060603. PMID: 40150453; PMCID: PMC11942098.
18. Wong K, Fayngersh A, Traba C, et al. (May 31, 2024) Using ChatGPT in the Development of Clinical Reasoning Cases: A Qualitative Study. Cureus 16(5): e61438. doi:10.7759/cureus.61438
19. Hadi A, Tran E, Nagarajan B, Kirpalani A (2024) Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. PLoS ONE 19(7): e0307383. https://doi.org/10.1371/journal.pone.0307383
20. Öncü, S., Torun, F. & Ülkü, H.H. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. BMC Med Educ 25, 278 (2025). https://doi.org/10.1186/s12909-025-06877-6
21. Hui, Z., Zewu, Z., Jiao, H. et al. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ 25, 50 (2025). https://doi.org/10.1186/s12909-024-06321-1
22. Thesen, T., Park, S.H. A generative AI teaching assistant for personalized learning in medical education. npj Digit. Med. 8, 627 (2025). https://doi.org/10.1038/s41746-025-02022-1
23. Simoni, J., Urtubia-Fernandez, J., Mengual, E. et al. Artificial intelligence in undergraduate medical education: an updated scoping review. BMC Med Educ 25, 1609 (2025). https://doi.org/10.1186/s12909-025-08188-2
24. Yang Y et al. Beyond Multiple-Choice Accuracy: Real-World Challenges of Implementing Large Language Models in Healthcare. Annual Review of Biomedical Data Science 2025. annualreviews.org https://doi.org/10.1146/annurev-biodatasci-103123-094851
25. Usuyama N, et al. Biomedical Natural Language Processing in the Era of Large Language Models. Annual Review of Biomedical Data Science. 2025. https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-071023-023244

26. Chan JTN, Kwek RK. Uncovering Bias and Variability in How Large Language Models Attribute Cardiovascular Risk. Frontiers in Digital Health. 2025. https://doi.org/10.3389/fdgth.2025.1508461
27. Jaleel A, et al. The Impact of ChatGPT on Academic Integrity in Medical Education. Frontiers in Education. 2024. https://doi.org/10.3389/feduc.2024.1356163
28. Ekaterina K, Ana M, Maia Z. Academic Integrity Within the Medical Curriculum in the Age of Generative Artificial Intelligence. Health Sci Rep. 2025 Feb 19;8(2):e70489. doi:10.1002/hsr2.70489. PMID: 39980818; PMCID: PMC11839754.
29. TEQSA. Gen AI – Academic Integrity and Assessment Reform. Gen AI Knowledge Hub. 2025. https://www.teqsa.gov.au/gen-ai-academic-integrity-and-assessment-reform
30. Natali C, et al. AI-Induced Deskilling in Medicine: A Mixed Method Literature Review for Setting a New Research Agenda. Artificial Intelligence Review. 2025. https://doi.org/10.1007/s10462-025-11352-1
31. Wong RSY. ChatGPT in Medical Education: Promoting Learning or Killing Critical Thinking? Education in Medicine Journal. 2024;16(2):177–183. https://doi.org/10.21315/eimj2024.16.2.13
32. Budzyń K, et al. *Endoscopist Deskilling Risk After Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study.* Lancet Gastroenterology & Hepatology. 2025.
33. Zhui L, Fenghe L, Xuehu W, Qining F, Wei R. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. J Med Internet Res. 2024 Aug 1;26:e60083. doi:10.2196/60083. PMID: 38971715; PMCID: PMC11327620.
34. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A (2023) ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2(2): e0000205. https://doi.org/10.1371/journal.pdig.0000205