RMMG
Revista Médica de Minas Gerais

REVIEW ARTICLE

# Use of Artificial Intelligences in undergraduate medical education: a scoping review

*Uso de Inteligências Artificiais na educação médica de graduação: uma revisão de escopo*

Davi Ricardo Soares Gama de Amorim[1], Daniel Gondim Malta[1], Abla Vitórya Nejaim Tenório Xavier[1], Rafael Vieira de Menezes[1], Hugo Rafael de Souza e Silva[1]

## ABSTRACT

**Introduction:** Artificial intelligence (AI) can personalize learning and enhance teaching in medical education, yet its use in undergraduate training is still little explored. This study mapped scientific production on AI in undergraduate medical teaching, identifying gaps and evidence for its integration. **Methods:** A scoping review was conducted in accordance with the items of the PRISMA Extension for Scoping Reviews, following five stages: (1) formulation of the research question using the Population, Concept, Context framework; (2) selection of databases and definition of search strategies; (3) export of retrieved studies to Rayyan and establishment of eligibility criteria; (4) selection of the retrieved articles by two blinded/independent reviewers; and (5) construction of a spreadsheet synthesizing the evidence, based on the Joanna Briggs Institute Manual for Evidence Synthesis. **Results:** The studies revealed innovative applications of AI in medical education. Technologies such as VOA, AmbuBPS, A-LRYNGO, and HipGuide supported practical training, enhancing student performance in specific procedures. COMET and machine learning–based tools showed results comparable to those of human tutors in assessment. Large language models, particularly ChatGPT, despite their limitations, showed competence in generating and solving questions across disciplines and producing medical texts. Interactive games with chatbots and virtual patients also demonstrated potential as learning support resources. **Conclusions:** The introduction of artificial intelligence in undergraduate medical education has proven beneficial for both teachers and students, with a range of applications spanning from enhancing learning dynamics to question formulation. However, further rigorous research is necessary to deepen understanding of the topic, generalize results, and facilitate the full integration of AI.

Keywords: Artificial Intelligence; Education, Medical, Undergraduate; Large Language Models

[1] Universidade de Pernambuco (UPE), Recife, Pernambuco, Brazil.

**Associate Editor Responsible:**
José Maria Peixoto
Faculdade de Medicina da UNIFENAS. Belo Horizonte, MG, Brazil.

**Corresponding author:**
Davi Ricardo Soares Gama de Amorim
Departamento de Medicina da UPE, Pernambuco, Brazil.
Email: davi.amorim@upe.br

**Supporting sources:**
None.

**Conflicts of interest:**
None.

Received on: September 10, 2025.
Approved on: February 12, 2026.
Publication Date: June 09 2026.

DOI: 10.5935/2238-3182.2026e36204-en

## RESUMO

**Introdução:** A inteligência artificial (IA) pode personalizar a aprendizagem e aprimorar o ensino em medicina, mas seu uso na graduação ainda é pouco explorado. Este estudo mapeou a produção científica sobre IA no ensino médico, identificando lacunas e evidências para sua integração. **Métodos:** Foi desenvolvida uma revisão de escopo, conforme os itens do PRISMA *Extension for Scoping Reviews*, que seguiu 5 etapas: 1) elaboração da pergunta de pesquisa seguindo o método População, Contexto, Conceito; 2) seleção das bases de dados e definição das estratégias de busca; 3) exportação dos estudos recuperados para o Rayyan e estabelecimento dos critérios de elegibilidade; 4) seleção dos artigos resgatados por 2 revisores blindados/independentes; 5) construção de uma planilha com a síntese das evidências, segundo o modelo *Joanna Briggs Institute Manual for Evidence Synthesis*. **Resultados:** Os estudos revelaram aplicações inovadoras da inteligência artificial na educação médica. Tecnologias como VOA, AmbuBPS, A-LRYNGO e HipGuide apoiaram o treinamento prático, melhorando o desempenho dos estudantes em procedimentos específicos. O COMET e ferramentas baseadas em aprendizado de máquina mostraram resultados comparáveis aos de tutores humanos na avaliação. Grandes modelos de linguagem, especialmente o ChatGPT, apesar das limitações, demonstraram competência em gerar e resolver questões em várias disciplinas e em produzir textos médicos. Jogos interativos com chatbots e pacientes virtuais também mostraram potencial como recursos de apoio ao aprendizado. **Conclusões:** A introdução da inteligência artificial no ensino de graduação em medicina tem se mostrado benéfica tanto para professores quanto para alunos, com uma gama de aplicações que vão desde a melhoria da dinâmica de aprendizagem até a formulação de perguntas. No entanto, é necessária uma investigação mais robusta para aprofundar a compreensão do tema, generalizar os resultados e facilitar a integração total da IA.

**Palavras-chave:** Inteligência Artificial; Educação de Graduação em Medicina; Modelos de Linguagem de Grande Escala;

## INTRODUCTION

Artificial Intelligence refers to a computational system that simulates and systematizes specific aspects of human intelligence or intelligent behavior, such as learning, intellectuality, and problem-solving[1]. This term was first employed in 1956 by John McCarthy at the Dartmouth Conference. According to him, artificial intelligence was "the science and engineering of making intelligent machines"[2].

In this sense, it is understood that AI is composed of three main paradigms: the symbolic (based on logic and knowledge), the statistical (probabilistic methods and machine learning), and the subsymbolic (embedded intelligence and research). These paradigms address various domains of problems, such as perception, reasoning, knowledge, planning, and communication[3]. Although there are different categories of artificial intelligence, machine learning (ML)[4], which is a system that trains a predictive model - identifying patterns in input data - and uses it to make useful predictions from new and previously unseen data [1], stands out.

However, it is worth noting that not every machine learning algorithm is universal and capable of solving all types of problems. Thus, it becomes crucial to select appropriate datasets to address a specific problem. In this regard,

Artificial Neural Networks play a significant role. They are mathematical models inspired by biological neural structures, acquiring computational ability through learning[5].

Another task of artificial intelligence involves natural language processing, which encompasses the machine's ability to understand human language. However, it does not simply consist of recognizing letters that form a word and then combining them with a definition; the algorithm is used to achieve an understanding of syntax and semantics to approximate the meaning of phrases, sentences, or paragraphs[4].

Due to its complexity, although artificial intelligence is often associated only with computers or robots, its application can be observed in various domains, spanning areas such as philosophy, psychology, linguistics, statistics[4], and even in undergraduate medical education.

Regarding medical education, the ability of AI to modify and personalize the learning process is one of the most evident future possibilities in this area. Artificial intelligence can contribute to meeting the specific needs of students amid the growing context of digitization, as they have distinct preferences and learning styles. AI can provide adaptive instructional content suitable for each student's knowledge gaps and learning speed by analyzing vast amounts of data and using machine learning algorithms[6].

However, the mentioned contribution of artificial intelligence in undergraduate medical education has not yet been fully elucidated in the literature, making it crucial for a deeper analysis of relevant information on the subject. For this reason, we conducted a scoping review to map scientific research on two axes: the use of artificial intelligences as teaching-learning support tools in undergraduate medical education and their proficiency in various curriculum components, allowing us to identify knowledge gaps and their evidence indicators, as well as to summarize and disseminate the main scientific findings.

## Objectives

To map the scientific literature on the use of artificial intelligences as support tools for teaching and learning in undergraduate medical education, as well as their proficiency in curriculum components, in order to identify knowledge gaps and evidence indicators.

## Methods

This study is a scoping review, whose development followed the PRISMA-ScR checklist[7]. The choice of scoping review as a research method is justified by its ability to identify, analyze, and investigate scientific evidence and gaps, as well as to understand the characteristics, concepts, and theoretical definitions of the proposed field of knowledge[8].

### Data source and search

The study was conducted in five stages: 1 - formulation of the research question following the Population (P), Concept (C), Context (C) framework - PCC; 2 - selection of databases to be used and definition of corresponding search terms and strategies; 3 - exportation of retrieved studies to the Rayyan platform and establishment of inclusion and exclusion criteria; 4 - selection of articles by 2 blinded/independent reviewers, who followed the criteria to ensure higher quality of evidence; 5 - construction of a spreadsheet synthesizing the evidence found, following the JBI Manual for Evidence Synthesis[9].

Thus, the research question was formulated following the PCC framework: 'To what extent does the adoption of artificial intelligences (AIs) as tools to support undergraduate medical education dynamize the pedagogical effectiveness of teachers and the learning experience of students?' P (population) – teachers and students of undergraduate medical courses; C (concept) – adoption of AIs to dynamize pedagogical effectiveness and learning experience; C (context) – undergraduate medical education.

To identify potentially relevant studies, we selected the following databases of scientific literature: PubMed, Scientific Electronic Library (SciELO), Web of Science via the main collection (Clarivate Analytics), Scopus (Elsevier), and Virtual Health Library (VHL). Next, search terms were defined based on the National Library of Medicine's (NLM) catalog, which compiles suitable MeSH Terms for the topic. These terms, associated with Boolean operators, formed the following search strategy: ("Artificial Intelligence" OR "AI" OR "Machine Learning" OR “Chatbot" OR "ChatGPT") AND ("Medical Education").

### Study selection and inclusion and exclusion criteria

Subsequently, the research retrieved from the five databases were exported on October 18, 2023, to the Rayyan reference manager, a tool developed by the Qatar Computing Research Institute (QCRI) to streamline manuscript selection. Simultaneously, the researchers established inclusion and exclusion criteria that guided the selection stage. Thus, manuscripts published between 2000 and 2023, in any language, originating from primary investigations addressing the application and/or analysis of artificial intelligence as support for teaching and learning in undergraduate medical education, were considered for inclusion in this study. Additionally, studies evaluating the proficiency of these tools in undergraduate curriculum components and in the performance of teaching activities were incorporated. Conversely, articles involving students from other fields, as well as theoretical studies, reviews, editorials, and opinion articles, were excluded due to their lower level of scientific evidence.

To reinforce the research's solidity, two blinded/independent assessors (D.R.S.G.A.; D.G.M.) followed a 4-step process to examine the retrieved publications: initially, duplicate articles between databases were removed to avoid analysis errors; then, Rayyan filters were used to exclude all review studies; subsequently, the titles and abstracts of the remaining studies were analyzed, excluding

those that did not meet the defined criteria; and finally, full-text reading of manuscripts whose titles and abstracts were inconclusive regarding eligibility criteria was conducted, including those that met the review's objective in this study. Any disagreements between the two assessors were resolved by a third researcher (R.V.M.), resulting in the set of articles composing this study. These selection steps and their respective quantities of associated articles were graphically organized in a flowchart (Figure 1).

### Data extraction and analysis

After completing the selection process, the two independent evaluators developed a data mapping form to identify the variables to be extracted from the reviewed studies. In an interactive process of form registration and updating, information related to the use of artificial intelligence as support for teaching and learning in undergraduate medical education was organized into a Word table, with any discrepancies being resolved by the senior researcher, H.R.S.S. The recorded variables followed the guidelines of the JBI Manual for Evidence Synthesis[9], covering aspects such as authorship, year of publication, country of origin, objectives, study design, sample size, and main outcomes, which composed the evidence board of the current study (Appendix 1), provided in the supplementary materials. Furthermore, the figures presented in this review were created using Microsoft Word.

## Results

During the publication identification phase, a total of 3,737 manuscripts were retrieved from various reputable databases. These sources included 1,491 from PubMed, 7 from SciELO, 348 from Web of Science, 1,203 from Scopus, and 688 from the Virtual Health Library. Subsequently, using the available filters in Rayyan, 1,174 duplicate studies were excluded, resulting in 2,563 citations. Of these, 2,442 manuscripts were removed, after analyzing titles and abstracts, leaving 121 articles. Among this remaining quantity, we excluded 72 for having study designs that did not align with the criteria of this review, 20 for not addressing the practical use of artificial intelligence in undergraduate medical education, and 5 for involving students from other courses. In the end, 24 articles remained eligible for inclusion in this review. The aforementioned screening process is presented in a flowchart (Figure 1).

In this scoping review, India stands out as the primary origin of the selected studies (n=5)[10-14], followed by the USA[15-17] and Canada[18-20], both with n=3 each. Subsequently, Germany[21,22] and South Korea[23,24] (n=2). Other represented countries include Australia[25], Saudi Arabia[26], Singapore[27], Denmark[28], France[29], Thailand[30], Taiwan[31], Turkey[32], and Serbia[33], each with 1 study.

Regarding the types of studies, more than half consist of Cross-Sectional Studies (n=13), which represents 54.16% of the total manuscripts manuscritos[10-16,18,21-23,26,32,33]. Among these, 92.30% are quantitative cross-sectional studies[10-14,16,18,21-23,26,32,33], and 7.69% are qualitative cross-sectional studies[15]. Next, Randomized Controlled Trials (n=5) are observed, representing 20.83% of the manuscripts[20,24,28,29,31]. Finally, Mixed Methods Studies (n=3) comprise 12.50% of the articles[17,25,30], and Cohort Studies (n=2) account for 8.33% of the citations[19,27].

Thus, the twenty-four studies investigated the use of Artificial Intelligences (AIs) as tools to support teaching and learning in undergraduate medical education. In order to facilitate understanding of the various applications of Artificial Intelligences in this context, the manuscripts were classified and organized into five categories. The 1st category encompasses studies where Artificial Intelligences were used in practical training of students, such as in surgical procedure simulations; the 2nd category gathers studies where AIs were employed in teaching activities, such as question generation; the 3rd category groups manuscripts that tested the proficiency of AIs in undergraduate medical curriculum components, such as solving questions from different disciplines; the 4th category includes research that evaluated the use of AIs in producing scientific information, compared to traditional sources of evidence; and finally, the 5th category focuses on studies that addressed the impact of these tools on students' knowledge acquisition and academic performance, through qualitative and quantitative analysis. To outline this classification, a table was created listing the articles corresponding to each category (Table 1).

Furthermore, it is worth noting that among the AI tools mentioned in the manuscripts, the ones that stood out the most were ChatGPT and its versions[10-15,18,21-23,26,32,33], being the subject of research in 54.16% of the studies. Following this, Microsoft Bing[10,12,21] and Google Bard[10,12,32] were cited in 12.50% of the articles each. The VOA (Virtual Operative Assistant) tool[19,20] and other machine learning (ML) models[17,25] were each mentioned in 8.33% of the references. Lastly, A-LRYNGO[24], AmbuBPS[28], COMET[30], ChatScript[16], Chatprogress[29], HipGuide[31], and VIP (Virtual Integrated Patient)[27] were individually referenced in 4.16% of the sources. The number of studies addressing each artificial intelligence can be seen in Figure 2, and the distribution of these AIs among the aforementioned categories is shown in Table 2.

### 1st category: use of artificial intelligences in practical training of students

Among the studies, 20.83% highlighted the use of AI in the practical training of students, considering parameters such as greater accuracy and technical quality, reduced execution time, and safety in performing procedures[19,20,24,28,31]. In this analysis, it was observed that the Virtual Operative Assistant (VOA) - an intelligent tutoring system - was able to enhance the accuracy and performance of students in the test group during tumor resection surgery, as its application was associated with a considerable reduction in the rate of removal of healthy brain tissue during surgery, indicating an improvement in procedure safety[19]. Another relevant aspect is that, following subpial resection cases, feedback

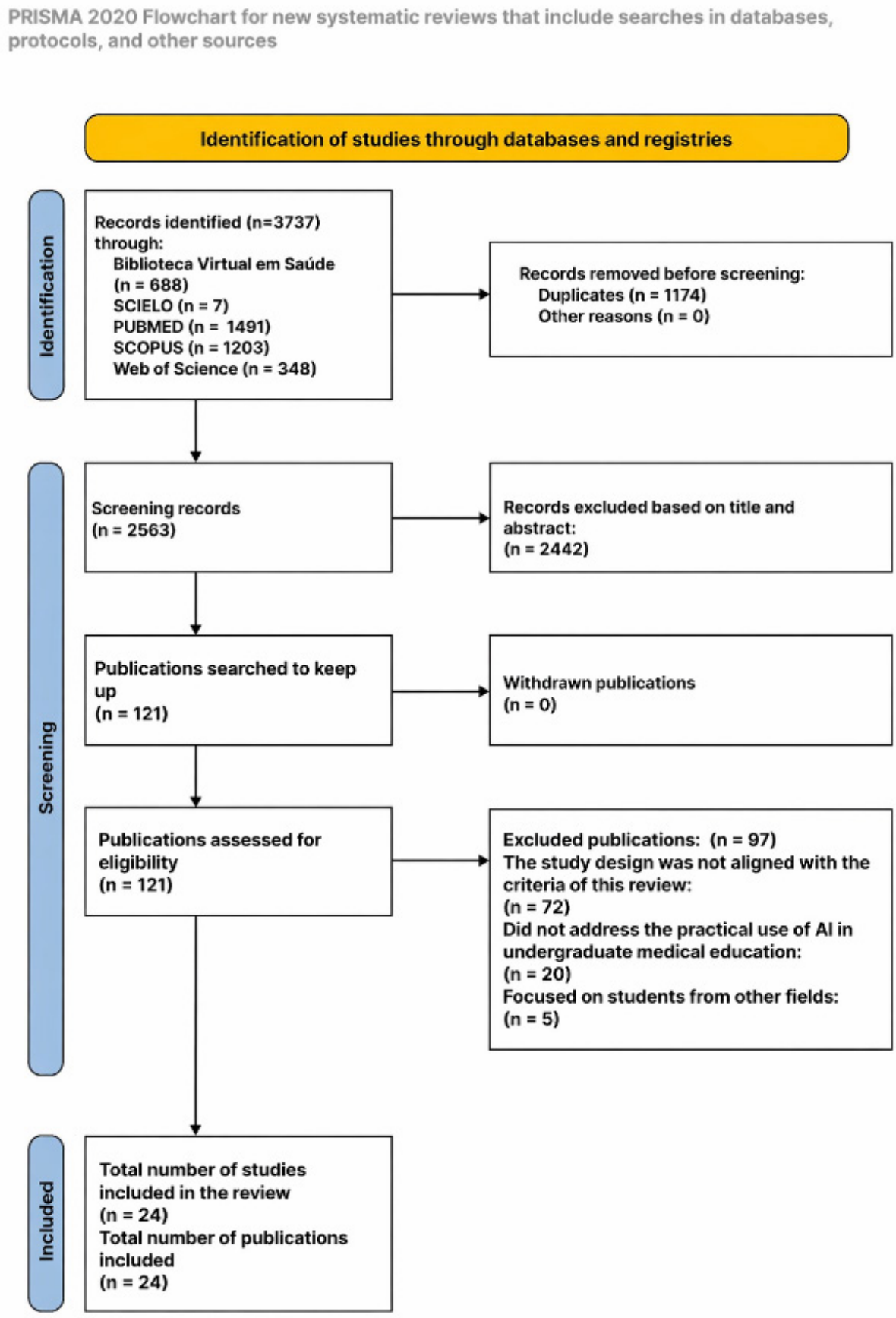


**Figure 1.** Identification of studies through databases and records (flowchart of manuscript selection).

**Source:** Elaborated by authors.

**Table 1.** Distribution of reviewed articles by analyzed category

| | |
|---|---|
| 1st Category: Use of AIs in practical training of students | 15, 16, 20, 24 and 27. |
| 2nd Category: Use of Artificial Intelligences in teaching activities | 8, 12, 13, 21, 26 and 28. |
| 3rd Category: Proficiency of Artificial Intelligences in undergraduate medical curriculum components | 6, 7, 9, 10, 14, 17, 18, 19, 22, 28 and 29. |
| 4th Category: Use of AIs in the production of scientific information compared to traditional evidence sources | 11 and 28. |
| 5th Category: Impact of AIs on students' knowledge acquisition and academic performance, through qualitative and quantitative analysis | 23, 24 and 25. |

**Source:** Elaborated by authors.

**Table 2.** Distribution of AIs by analyzed category.

| | |
|---|---|
| 1st Category: Use of AIs in practical training of students | Virtual Operative Assistant (VOA), AmbuBPS, A-LRYNGO, HipGuide. |
| 2nd Category: Use of Artificial Intelligences in teaching activities | COMET, ML, ChatScript, ChatGPT, Microsoft Bing and Google Bard. |
| 3rd Category: Proficiency of Artificial Intelligences in undergraduate medical curriculum components | ChatGPT, ChatGPT-4, ChatGPT-3.5-Turbo, Microsoft Bing and Google Bard. |
| 4th Category: Use of AIs in the production of scientific information compared to traditional evidence sources | ChatGPT and Google Bard. |
| 5th Category: Impact of AIs on students' knowledge acquisition and academic performance, through qualitative and quantitative analysis | Chatprogress, Virtual Integrated Pacient (VIP) and AmbuBPS. |

**Source:** Elaborated by authors.

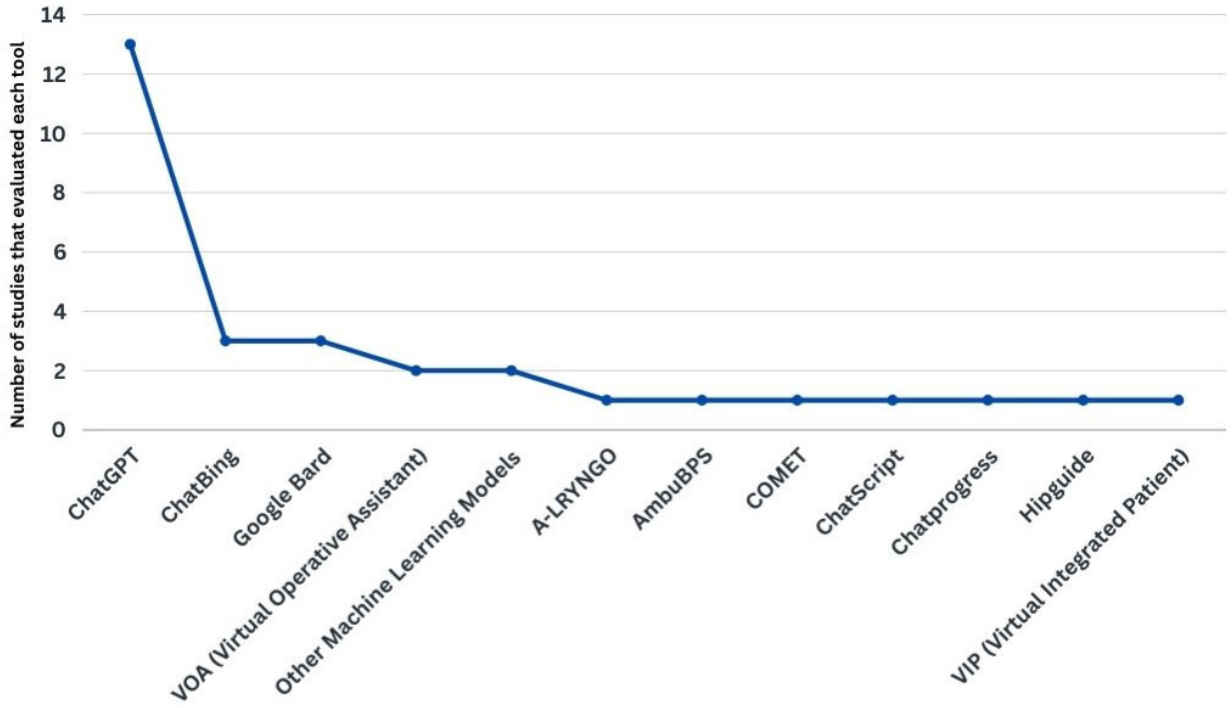


**Figure 2.** Number of studies that analyzed, individually or in combination, each Artificial Intelligence tool addressed in the present review.

**Source:** Elaborated by authors.

from artificial intelligence (VOA) may be more effective in improving performance in simulated surgical procedures than instruction by experts or the absence of feedback[20]. However, it is worth noting that some unintended changes occurred, such as a decrease in the speed and acceleration of the dominant hand, which may require further evaluation[19].

Similarly, other artificial intelligences, such as AmbuBPS, A-LRYNGO, and HipGuide, also yielded positive results in improving learning and practical training in bronchoscopies, laryngoscopies, and fracture analysis, respectively[24,28,31]. AmbuBPS is a system that utilizes artificial intelligence and is designed to assist in medical training. The technology includes a single-use scope and a set of real-time features, such as on-screen labels, a diagram of the bronchial tree, and a Structured Progression (SP) score. These features aim to enhance bronchoscopy practice by providing real-time feedback on positioning and recognition of bronchial segments. In the study in question, AmbuBPS provided superior results compared to the control group, especially in dynamic coordination, trial time, and pattern sequence analysis, allowing students to feel more confident with the procedure and more engaged in bronchoscopy learning[28]. In the context of laryngoscopy technique, glottic orientation based on artificial intelligence using A-LRYNGO - a channel-type videolaryngoscope coupled with an artificial intelligence system - was promising, achieving accurate orientation in most participants. The study highlighted that practical training contributed to the improvement of students' performance, demonstrating noticeable differences among various laryngoscopes and emphasizing the positive influence of artificial intelligence on glottic orientation[24]. Regarding the use of the HipGuide, an AI system designed for hip fracture detection, in orthopedics, participants were divided into two groups: AIL (test group) and CL (control group). In the CL group, post-learning scores showed no significant differences compared to pre-learning, whereas in the AIL group, all students recorded significant improvements with AI-assisted interpretation and analysis of hip fractures, resulting in scores considerably higher than those obtained in pre-learning[31].

## 2ND CATEGORY: USE OF ARTIFICIAL INTELLIGENCES IN TEACHING ACTIVITIES

Another aspect analyzed was the use of artificial intelligences in carrying out teaching activities, such as tutoring, question development, or even for student activity assessment[12,16,17,25,30,32], totaling 25% of the studies. In this parameter, it can be seen that AIs like COMET and ML tools managed to match or even surpass human tutors and faculty in student assessment in specific activities[17,30]. The COMET intelligent tutoring system demonstrated effectiveness comparable to that of human tutors for monitoring and evaluating medical students undergoing the PBL teaching method[30]. Similarly, machine learning (ML) proved promising in the formative analysis of clinical reasoning written by medical students, being able to assess diagnostic justification essays as reliably as, or more reliably than, faculty evaluations[17]. On the other hand, ChatScript did not exhibit the same efficacy in this type of task, performing worse than humans in evaluating medical students' performance in Objective Structured Clinical Examinations (OSCEs), tests aimed at assessing communicative skills in the context of clinical conversations and medical interviews, as observed in another study[16]. Furthermore, in qualitative aspect, there was divergence among students regarding computer-based marking (Machine Learning) of written exams. In the mentioned research, more than half of the students expressed doubts about the ability of computers to perform marking accurately enough, although most agreed that computer marking would be more objective than human marking[25].

Regarding the ability to formulate questions, ChatGPT, Microsoft Bing and Google Bard demonstrated potential in performing this task but also exhibited limitations[12,32]. One of the reviewed studies examined the ability of these tools to create multiple-choice questions on physiology and highlighted the success of ChatGPT and Google Bard in all assessed content competencies, differing from Bing, which showed shortcomings in two competencies. However, it is important to note that none of the artificial intelligences were able to generate a significant number of multiple-choice questions requiring a high level of comprehension[12] of the subject. In another study evaluating different versions of ChatGPT and Google Bard in creating anatomy questions, some of the questions generated by ChatGPT were considered incorrect by anatomist evaluators[32].

## 3RD CATEGORY: PROFICIENCY OF ARTIFICIAL INTELLIGENCES IN UNDERGRADUATE MEDICAL CURRICULAR COMPONENTS

Additionally, we sought to understand the proficiency of AIs in the curricular components of undergraduate medical education, both in question-solving and case studies. From this perspective, 41.67% of the studies can be grouped in this axis of analysis[10,11,13,14,18,21-23,26,32,33]. In this context, it is worth highlighting ChatGPT, which stood out as the most effective artificial intelligence in dealing with questions from various disciplines such as anatomy and physiology, besides demonstrating its utility in applying physiology case studies and in biostatistical analyses. In order to illustrate the distribution of proficiency studies in various areas of undergraduate education, a graph was created (Figure 3), and their respective references were organized through Table 3.

In one of the studies, the analysis focused on the performance of three Large Language Models (LLMs): GPT-4, GPT-3.5-Turbo and Bing. Specifically, the abilities of these models to solve questions taken from the 2022 German Medical State Exams were examined, considering scenarios with and without media presence in the questions. The performances of the AIs were compared among themselves and in relation to students in the spring and fall 2022 exams. Comparing the artificial intelligences among themselves in the spring exam, GPT-4 stood out by showing better performance in solving questions with media, followed by Bing and GPT-3.5-Turbo. When

questions with media were excluded, Bing showed better performance in solving questions, followed by GPT-4 and GPT-3.5-Turbo. In comparison to the students, GPT-4 and Bing showed superior performance, while GPT-3.5-Turbo showed comparable performance to the students. Comparing the artificial intelligences among themselves in the fall exam, GPT-4 demonstrated better performance in solving questions with media, followed by Bing and GPT-3.5-Turbo. When questions with media were removed, Bing and GPT-4 showed similar performances, while GPT-3.5-Turbo had inferior performance. In comparison to the students, GPT-4 and Bing maintained superior performance in solving questions, but GPT-3.5-Turbo obtained lower results than the students in this second exam. Overall, GPT-4 showed the best results, followed by Bing with intermediate performance, and lastly, GPT-3.5-Turbo with the lowest performance[21]. In another manuscript, the proficiency of artificial intelligences in solving anatomy questions in different linguistic domains was compared. With this purpose, the responses provided by the AIs to two distinct users who asked the same questions were analyzed. The intention was to verify if the answers obtained in separate accesses were consistent or exhibited variations. Thus, it was revealed that ChatGPT in Turkish language achieved the highest performance, followed by Google Bard, also in Turkish. The other AIs, including the first user of Google Bard in Turkish, ChatGPT in English, and Google Bard in English, showed varied rates. However, none of them reached a significant score[32].

Another article evaluated the performance of AIs in physiology case studies, demonstrating that ChatGPT consistently showed the best performance, obtaining higher scores - attributed by both evaluators - compared to Bing and Google Bard. Detailed data reveals that, on average, ChatGPT achieved a higher performance score among the three AIs, considering different physiological systems, while Bing recorded significantly lower scores. Bard positioned itself as intermediate in this evaluation[10].

Regarding the performance of Artificial Intelligences in solving questions from curricular components, it is crucial to note the divergent results presented by one manuscript, especially concerning the ability to solve biostatistical problems in different versions of ChatGPT. GPT-3.5 demonstrated efficacy in solving tasks involving categorical data, cross-sectional study, and paired t-test, however, it faced difficulties in questions related to chi-square test, one-way analysis of variance (ANOVA), and sample size. In contrast, GPT-4 had a more comprehensive performance, managing to solve some of these additional questions in subsequent attempts[33].

Additionally, it is pertinent to examine the performance of ChatGPT responses concerning the content of the questions. In a study involving multiple-choice questions,

**Table 3.** Analyzed curriculum components and their respective studies

| | |
|---|---|
| Multidisciplinary | 17, 18, 22, 14. |
| Physiology | 6, 9. |
| Anatomy | 28. |
| Biostatistics | 29. |
| Microbiology | 7. |
| Biochemistry | 10. |
| Parasitology | 19. |

**Source:** Elaborated by authors.

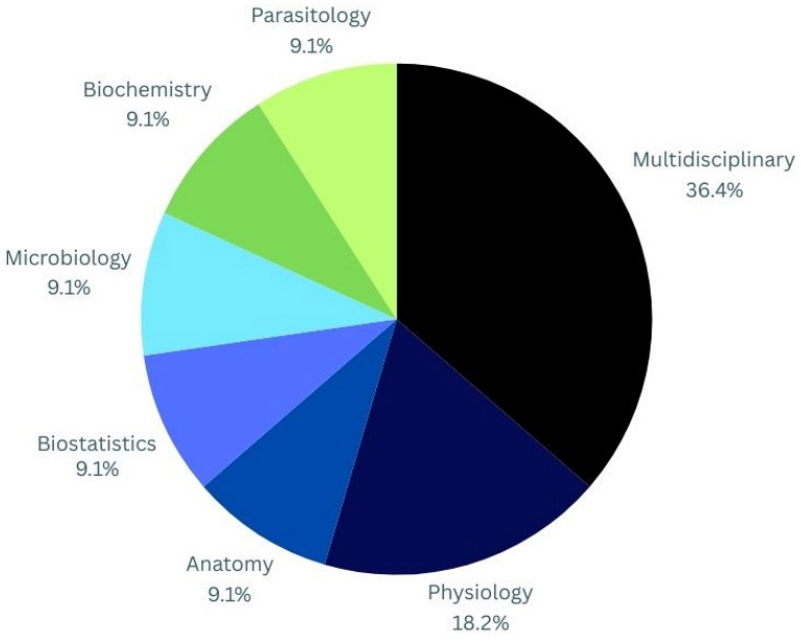


**Figure 3.** Professional background of proficiency studies by analyzed curricular component.

**Source:** Elaborated by authors.

it was observed that ChatGPT scored lower in forensic medicine and radiology, contrasting with a high score in dermatology[22]. Similarly, when evaluating the effectiveness of ChatGPT in solving fundamental physiology questions, it stands out that cardiovascular physiology achieved the highest score, followed by neurophysiology, while endocrine physiology recorded the lowest score[13]. In a study on ChatGPT's ability to answer microbiology questions, it was found that the category of central nervous system infections obtained the highest score, while gastrointestinal and hepatobiliary systems obtained the lowest score[11]. Reinforcing these analyses, it was found that ChatGPT demonstrated the ability to answer higher-order questions in the field of medical biochemistry, although its performance varied among the competencies of the discipline[14]. Lastly, in another study conducted with ChatGPT, it was noted that it was able to answer multiple-choice questions in basic and clinical medical sciences, demonstrating a satisfactory level of knowledge. However, the performance was slightly higher in basic questions compared to clinical questions[26].

On the other hand, in a study comparing students' responses to ChatGPT's responses in solving Concept Application Exercises (CAEs), the scores assigned by tutors revealed a superior quality of ChatGPT's responses compared to those of the student participants. However, both ChatGPT's performance and that of the students in the research were inferior compared to a group of students previously subjected to CAEs[18]. Similarly, in an analysis on parasitology, ChatGPT's performance was considered inferior to that of medical students. Some limitations identified in this Artificial Intelligence included the inability to interpret figures and graphs, unfamiliarity with specific epidemiological data from Korea, and difficulty in understanding multiple-choice questions[23].

### 4TH CATEGORY: USE OF ARTIFICIAL INTELLIGENCES IN THE PRODUCTION OF SCIENTIFIC INFORMATION COMPARED TO TRADITIONAL EVIDENCE SOURCES

Additionally, the quality of scientific productions, especially articles, generated by artificial intelligences, was evaluated in comparison to traditional sources of evidence, accounting for 4.17% of the studies[15,32]. One study investigated the perception of medical students and laypersons regarding articles produced by ChatGPT, contrasted with information from an evidence-based surgery book. This research revealed that medical students considered the book texts more comprehensive than those generated by ChatGPT in addressing each of the 5 chosen surgical conditions. However, ChatGPT articles were considered clearer by medical students in 4 of the studied surgical pathologies. Simultaneously, laypersons from the general population noticed superior clarity in ChatGPT articles related to appendicitis and better organization in those addressing diverticulitis[15]. In another manuscript in this category, the scientific writing ability on anatomy by ChatGPT and Google Bard was evaluated. It was found that articles produced by both AI were less detailed and of a basic level compared to previously prepared scientific works on the subject, and they lacked references. However, it is worth noting that among the two analyzed artificial intelligences, ChatGPT stood out in creating text more adherent to academic standards, presenting a more refined structure aligned with the standard organization required for this type of content[32].

### 5TH CATEGORY: IMPACT OF ARTIFICIAL INTELLIGENCES ON KNOWLEDGE ACQUISITION AND ACADEMIC PERFORMANCE OF STUDENTS, THROUGH A QUALITATIVE AND QUANTITATIVE ANALYSIS

Furthermore, through both qualitative and quantitative approaches, the impact on knowledge acquisition and academic performance of students using AIs was measured, accounting for 12.50% of the studies[27-29]. One of these studies emphasized the positive impact of interactive games incorporating artificial intelligence, utilizing Chatbot technology (Chatprogress), in addressing topics in pneumology. This educational methodology not only exceeded expectations by enhancing students' performance in pneumology subtests but also provided benefits for the overall examination, which covered areas such as cardiology, intensive care, and pneumology. Additionally, qualitative data revealed high student satisfaction with the new learning tool, demonstrating continued interest in its use[29]. Another study showed that the use of the Virtual Integrated Patient (VIP) - a virtual patient simulator leveraging advances in artificial intelligence in the field of naturalistic conversation technology - helped medical students remember the methodological content taught in classes, improved understanding of different clinical cases, increased confidence, and enhanced their performances in conducting patient histories[27]. Similarly, the application of AmbuBPS in performing bronchoscopies was well received by students, who reported confidence in performance during the procedure associated with the use of the technology, recommended its use, and showed continued interest in incorporating this tool into their learning[28].

## 5. DISCUSSION

After analysis, the evidence suggests that Artificial Intelligence presents significant potential to enhance undergraduate medical education. The versatility of this technology offers a wide range of applications in the university context, serving as an effective tool to support the development of medical students and optimize educators' performance. In 87.50% of the studies, its use in the educational environment proved to be useful and beneficial in some way, contrasting with only 12.5% of the studies in which AI was unable to fulfill the assigned tasks or demonstrated unsatisfactory performance.

### 1ST CATEGORY: USE OF ARTIFICIAL INTELLIGENCES IN PRACTICAL TRAINING OF STUDENTS

Practical training in undergraduate medical education is fundamental for professional development, as it enables students to acquire the necessary skills for performing

procedures (surgical or otherwise), familiarity with medical equipment and application of theoretical knowledge from early on in their education. Thus, considering the use of artificial intelligence in developing such practical skills by students, it is evident that these technologies hold promise for enhancing practical techniques and ensuring safety in their execution. This was evidenced based on the articles presented, through the use of artificial intelligences such as VOA[19,20], AmbuBPS[28], A-LRYNGO[24] and HipGuide[31], which facilitated faster and higher-quality skill acquisition by students. Consequently, these results provided students with greater confidence in reproducing such procedures later on, thereby minimizing the risk of errors in professional practice. In this context, this review aligns with the studies of Chan et al[3] and Narayanan et al[34]., as our findings corroborate the advantages and utilities highlighted by them regarding adaptive instructional systems. These technologies capable of mimicking human tutors benefit students' learning process, underscoring the importance of incorporating artificial intelligences into students' practical training from early education onwards.

### 2ND CATEGORY: USE OF ARTIFICIAL INTELLIGENCES IN TEACHING ACTIVITIES

Additionally, the analysis of artificial intelligence in performing teaching activities proved to be relevant. Considering the challenges faced by university professors, such as intense workload and multiple associated responsibilities including activity grading, question development, and tutoring, the use of AI as an aid in performing these tasks becomes a promising alternative. In this context, artificial intelligence-based technologies like the intelligent tutoring system (COMET) and some machine learning models (ML) can be adopted as tools to facilitate the assessment of student activities. This is evidenced by the data collected in this research, which indicate that these AIs demonstrated performance equivalent to or even superior to humans in this task. This artificial intelligence approach thus strengthens the applications highlighted by other studies available in the scientific literature[3,34], which previously emphasized the ability of automated essay scoring (AES) systems to analyze student work, identify study patterns and concepts, and provide individualized feedback.

Furthermore, chatbots, including ChatGPT, despite still having limitations that require continuous improvement of these AIs, have shown potential as assisting tools for teachers in various tasks, such as question development, as indicated in studies included in this research.

However, there are essential considerations to be weighed before officially incorporating any AI as a teaching and learning tool. One of them involves the need for careful screening to ensure that the AI used is the most appropriate and reliable. As evidenced in this research, ChatScript did not demonstrate sufficient competence in assessing the performance of medical students in Objective Structured Clinical Examinations (OSCEs), highlighting that not all AIs are suitable for effectively performing teaching activities. The other consideration is related to the acceptance of the use of these tools by the student body, as not all students fully trust their accuracy.

### 3RD CATEGORY: PROFICIENCY OF ARTIFICIAL INTELLIGENCES IN UNDERGRADUATE MEDICAL CURRICULAR COMPONENTS

Furthermore, it was observed that the ability of artificial intelligences to solve questions and case studies in various components of the medical undergraduate curriculum is promising. Based on the results of the articles used in our research, we noticed that AIs, such as ChatGPT, for example, can be used as learning enhancement tools. Solving questions enables us to apply theoretical study, identify knowledge gaps, prepare for exams, and develop clinical reasoning. In this context, ChatGPT stood out for both solving questions and case studies in some disciplines. Additionally, other versions of ChatGPT and Bing were also used as preparation tools for medical exams and proved to be effective study tools for solving questions and better preparation by students.

However, it was evidenced that questions with information contained in media are not processed or interpreted by artificial intelligences. Furthermore, exercises involving some curriculum components, such as forensic medicine and radiology, did not yield such satisfactory results, and one study pointed out that students' responses were superior to those of ChatGPT in another subject. Thus, it became clear that the proficiency of artificial intelligences in curriculum components still has limitations.

### 4TH CATEGORY: USE OF ARTIFICIAL INTELLIGENCES IN THE PRODUCTION OF SCIENTIFIC INFORMATION COMPARED TO TRADITIONAL EVIDENCE SOURCES

Regarding the quality of scientific information generated by artificial intelligences, it is observed that, although the materials produced by them have a certain level of quality, they still cannot effectively replace traditional sources of evidence, such as books and conventional articles. Despite the remarkable ability of some AIs, such as ChatGPT and Google Bard, to generate easily understandable articles, their texts often lack depth and comprehensiveness compared to those found in traditional sources. Thus, although these AIs can be used as quick search tools to assist medical students in their studies during undergraduate education, their use, as of the time of this study, does not fully replace the use of traditional scientific evidence sources as the most appropriate place for accessing information.

### 5TH CATEGORY: IMPACT OF ARTIFICIAL INTELLIGENCES ON KNOWLEDGE ACQUISITION AND ACADEMIC PERFORMANCE OF STUDENTS, THROUGH A QUALITATIVE AND QUANTITATIVE ANALYSIS

Finally, it became clear that, in addition to quantitative aspects regarding the impacts of using artificial intelligence in various aspects of medical education, the qualitative effects, obtained from forms or questionnaires, were also important for understanding the impact of AI use on students'

learning and academic performance. The combination of quantitative and qualitative data revealed not only tangible improvements in learning, such as simulating realistic situations and active engagement through games (as in the Chatbot), but also emphasized students' confidence in their performance associated with the technologies. These benefits associated with the use of Artificial Intelligence are also in line with the advantages highlighted in the study by Narayanan et al. (2023)[34], which emphasizes the usefulness and various applications of these technologies as support tools for undergraduate student learning.

Furthermore, the ongoing interest of students in the integration of these tools, as demonstrated in the articles included in this research, suggested that artificial intelligences not only enhanced learning but also became recommended and desired elements in medical education. Thus, this convergence of results strengthened the positive outlook regarding the integration of AI, indicating a favorable impact on students' learning and academic performance in the field of medicine.

### Limitations

This scoping review presents several limitations to be considered. A recurring limitation is the focus on a specific set of large language models (LLMs), such as ChatGPT, Bard, and Bing, which may result in a limited view of the diversity available in this field. The assessment of the performance of these LLMs often relies on responses to predefined cases, limiting generalization to a wider variety of clinical scenarios, with additional concerns about bias due to training data. The subjectivity introduced by expert evaluations, the lack of generalization to other medical specialties, and the absence of considerations about potential biases in LLMs highlight the complexity of the results. Critical issues include the inability of some systems, such as ChatGPT, to interpret complex interactions and the limitation in inputting information, such as charts and figures. The study also did not fully address potential biases or intrinsic errors of LLMs, pointing to the need for caution in interpreting the results. Several studies presented methodological challenges, such as the limitation in the number of investigated biostatistical problems and the lack of pre-established guidelines for interactions with ChatGPT, underscoring the demand for more comprehensive and refined research. The review highlights that, while these studies provide valuable insights, it is essential to recognize and address these limitations for a more complete and reliable understanding of the role of LLMs in medical education and other areas. Additionally, some studies presented limitations related to the observed sampling that may compromise the generalizability and applicability of the results.

## Conclusion

The aim of this scoping review was to map the scientific literature on the use of artificial intelligence as tools to support teaching and learning in undergraduate medical education, in order to identify knowledge gaps and evidence clues. In this regard, we identified five main areas of AI application: 1 - using artificial intelligence in practical training of students; 2 - employing artificial intelligence in teaching activities; 3 - assessing the proficiency of artificial intelligence in undergraduate medical curriculum components; 4 - utilizing artificial intelligence in scientific information production; 5 - examining the impact of AIs on students' knowledge and academic performance. The results of our study suggest that the introduction of artificial intelligence in the academic environment is beneficial for both teachers and undergraduate students, allowing for the dynamization of the development of materials and tutorials, improvement in precision, technical quality, and safety of procedures, as well as favoring knowledge acquisition and problem-solving. However, it is worth noting some identified limitations, such as the need for improvement in question development, challenges in interpreting graphics by certain artificial intelligences, and insufficient depth of the produced scientific information. Therefore, this review highlights the pressing need for comprehensive and refined research to achieve a more detailed understanding of the role of artificial intelligence in undergraduate medical education. The pursuit of a more robust methodological approach in subsequent studies will be crucial to validate and generalize the conclusions, thus contributing to the effective advancement of this area of study.

## Authors' Contributions

The authors' contributions are structured according to the CRediT taxonomy, as described below: Conceptualization & Project Administration: DRSG de Amorim. Data Curation & Formal Analysis: AVNT Xavier; RV de Menezes. Investigation & Methodology: DG Malta; DRSG de Amorim. Supervision & Validation: HR de S e Silva. Visualization: DG Malta. Writing – Original Draft: AVNT Xavier; DG Malta; DRSG de Amorim; RV de Menezes. Writing – Review & Editing: AVNT Xavier; DG Malta; DG Malta; DRSG de Amorim; RF de Menezes.

## Copyright

Copyright© 2024 Amorim et al. This is an Open Access article distributed under the terms of the *Creative Commons* Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original article is properly cited.

## References

1. Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. Healthc Manage Forum. 2020;33(1):10-8. DOI: https://doi.org/10.1177/0840470419873123.

2. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem Rev. 2019;119(18):10520-94. DOI: https://doi.org/10.1021/acs.chemrev.8b00728.

3. Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. JMIR Med Educ. 2019;5(1):e13930. DOI: https://doi.org/10.2196/13930.

4. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology. 2020;132(2):379-94. DOI: https://doi.org/10.1097/ALN.0000000000002960.

5. Ludemir TB. Artificial Intelligence and Machine Learning: Current State and Trends. Estud Av. 2021;35(101):85–94. DOI: https://doi.org/10.1590/s0103-4014.2021.35101.007.

6. Mir MM, Mir GM, Raina NT, Mir SM, Mir SM, Miskeen E, et al. Application of Artificial Intelligence in Medical Education: Current Scenario and Future Perspectives. J Adv Med Educ Prof. 2023;11(3):133-40. DOI: https://doi.org/10.30476/JAMP.2023.98655.1803.

7. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanations. Ann Interm Med. 2018; 169(7):467–73. DOI: https://doi.org/10.7326/m18-085.

8. Mattos SM, Cestari VRF, Moreira TMM. Scoping protocol review: PRISMA-ScR guide refinement. Rev Enferm UFPI. 2023;12(1). DOI: https://doi.org/10.26694/reufpi.v12i1.3062.

9. Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI. 2020. DOI: https://doi.org/10.46658/JBIMES-20-01.

10. Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. Cureus. 2023;15(8):e42972. DOI: https://doi.org/10.7759/cureus.42972.

11. Das D, Kumar N, Longjam LA, Sinha R, Roy AD, Mondal H, et al. Assessing the Capability of ChatGPT in Answering First- and Second-Order Knowledge Questions on Microbiology as per Competency-Based Medical Education Curriculum. Cureus. 2023;15(3):e36034. DOI: https://doi.org/10.7759/cureus.36034.

12. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023;15(6):e40977. DOI: https://doi.org/10.7759/cureus.40977.

13. Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the Efficacy of ChatGPT in Solving Questions Based on the Core Concepts in Physiology. Cureus. 2023;15(8):e43314. DOI: https://doi.org/10.7759/cureus.43314.

14. Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. Cureus. 2023;15(4):e37023. DOI: https://doi.org/10.7759/cureus.37023.

15. Breeding T, Martinez B, Patel H, Nasef H, Arif H, Nakayama D, et al. The Utilization of ChatGPT in Reshaping Future Medical Education and Learning Perspectives: A Curse or a Blessing? Am Surg. 2024;90(4):560-6. DOI: https://doi.org/10.1177/00031348231180950.

16. Maicher KR, Zimmerman L, Wilcox B, Liston B, Cronau H, Macerollo A, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. Med Teach. 2019;41(9):1053-9. DOI: https://doi.org/10.1080/0142159X.2019.1616683.

17. Cianciolo AT, LaVoie N, Parker J. Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence. Acad Med. 2021;96(7):1026-35. DOI: https://doi.org/10.1097/ACM.0000000000004010.

18. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an Undergraduate Medical Program. J Med Educ Curric Dev. 2023;10:23821205231204178. DOI: https://doi.org/10.1177/23821205231204178.

19. 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. DOI: https://doi.org/10.1001/jamanetworkopen.2023.34658.

20. 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. DOI: https://doi.org/10.1001/jamanetworkopen.2021.49008.

21. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany. JMIR Med Educ. 2023;9:e46482. DOI: https://doi.org/10.2196/46482.

22. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online. 2023;28(1):2220920. DOI: https://doi.org/10.1080/10872981.2023.2220920.

23. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study. J Educ Eval Health Prof. 2023;20:1. DOI: https://doi.org/10.3352/jeehp.2023.20.1.

24. Choi J, Lee Y, Kang GH, Jang YS, Kim W, Choi HY, et al. Educational suitability of new channel-type video-laryngoscope with AI-based glottis guidance system for novices wearing personal-protective-equipment. Medicine (Baltimore). 2022;101(9):e28890. DOI: https://doi.org/10.1097/MD.0000000000028890.

25. Hollis-Sando L, Pugh C, Franke K, Zerner T, Tan Y, Carneiro G, et al. Deep learning in the marking of medical student short answer question examinations: Student perceptions and pilot accuracy assessment. Focus Health Prof Educ Multi-Prof J. 2023;24(1):38–48. DOI: https://doi.org/10.11157/fohpe.v24i1.531.

26. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthcare (Basel). 2023;11(14):2046. DOI: https://doi.org/10.3390/healthcare11142046.

27. Kong JSM, Teo BS, Lee YJ, Pabba AB, Lee EJD, Snh JCG. Virtual Integrated Patient: An AI supplementary tool for second-year medical students. Asia Pac Sch. 2021;6(3):87–90. DOI: https://doi.org/10.29060/taps.2021-6-3/sc2394.

28. 28. Cold KM, Xie S, Nielsen AO, Clementsen PF, Konge L. Artificial Intelligence Improves Novices' Bronchoscopy Performance: A Randomized Controlled Trial in a Simulated Setting. Chest. 2024;165(2):405-13. DOI: https://doi.org/10.1016/j.chest.2023.08.015.

29. 29. Al Kahf S, Roux B, Clerc S, Bassehila M, Lecomte A, Moncomble E, et al. Chatbot-based serious games: A useful tool for training medical students? A randomized controlled trial. PLoS One. 2023;18(3):e0278673. DOI: https://doi.org/10.1371/journal.pone.0278673.

30. Suebnukarn S, Haddawy P. A Bayesian approach to generating tutorial hints in a collaborative medical problem-based learning system. Artif Intell Med. 2006;38(1):5-24. DOI: https://doi.org/10.1016/j.artmed.2005.04.003.

31. Cheng CT, Chen CC, Fu CY, Chaou CH, Wu YT, Hsu CP, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. Insights Imaging. 2020;11(1):119. DOI: https://doi.org/10.1186/s13244-020-00932-0.

32. Ilgaz HB, Çelik Z. The Significance of Artificial Intelligence Platforms in Anatomy Education: An Experience With ChatGPT and Google Bard. Cureus. 2023;15(9):e45301. DOI: https://doi.org/10.7759/cureus.45301.

33. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. J Educ Eval Health Prof. 2023;20:28. DOI: https://doi.org/10.3352/jeehp.2023.20.28.

34. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial Intelligence Revolutionizing the Field of Medical Education. Cureus. 2023;15(11):e49604. DOI: https://doi.org/10.7759/cureus.49604.

This is an open access article distributed under the terms of the Creative Commons Attribution License.

## Appendix 1. Evidence synthesis.

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Suebnukarn, Haddawy. | 2006 | Thailand | 120 medical students and 10 tutors | Mixed methods study | Development of representational techniques and algorithms for generating tutoring hints in group problem-solving in PBL, as well as the implementation of these techniques in a collaborative intelligent tutoring system, COMET. | On average, 74.17% of human tutors used the same hint as COMET. The most similar human tutor agreed with COMET 83% of the time, and the least similar tutor agreed 62% of the time. Our results show that COMET hints agree with the hints of most human tutors with a high degree of statistical agreement (McNemar test, $p$=0.652, $\kappa$ = 0.773). The focus of the group activity chosen by COMET agrees with that chosen by most human tutors with a high degree of statistical agreement (McNemar test, $p$=0.774, $\kappa$ = 0.823). Bayesian network clinical reasoning models can be combined with generic tutoring strategies to successfully emulate human tutor hints in medical group PBL. |
| Maicher et al. | 2019 | USA | Nonexistent | Quantitative cross-sectional study | To describe a standardized virtual patient system (VPS) that allows students to practice their history-taking skills and receive immediate feedback. | The overall scores computed by the computer were not different from those provided by human assessors, and the overall accuracy of the computer system was 87%, compared to 90% for human assessors. Inter-rater reliability was high in 19 out of 21 categories. We have developed a standardized virtual patient system that can understand, respond to, categorize, and assess students' performance in information gathering during a typical medical history, thereby allowing students to practice their history-taking skills and receive immediate feedback. |
| Cheng et al. | 2020 | Taiwan | 30 medical students | Randomized controlled trial | To apply a prospective blind image education system developed by them and confirming if AI can support medical students and help them learn with confidence. | Pre-learning performance was comparable in both groups. In the CL group, post-learning accuracy (78.66 ± 14.53) was higher than pre-learning accuracy (75.86 ± 11.36) without significant difference ($p$=0.64). The AIL group showed notable improvement. The WithAI score (88.87 ± 5.51) was significantly higher than the pre-learning score (75.73 ± 10.58, $p < 0.01$). Furthermore, post-learning score (84.93 ± 14.53) was better than pre-learning score ($p < 0.01$). The increase in accuracy was significantly greater in the AIL group than in the CL group. The study demonstrated the feasibility of AI to enhance medical education. Integration of AI into medical education requires dynamic collaboration from research, clinical, and educational perspectives. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Cianciolo, LaVoie, Parker | 2021 | USA | 121 e medical students e 52 clinical professors | Mixed methods study | To assess the attitudes of medical students and faculty towards AI, in preparation for teaching AI fundamentals and data science applications in clinical practice within an integrated medical education curriculum. | Only 30% of students and 50% of faculty responded that they were aware of AI topics in medicine. The majority of students (72%) and faculty (59%) learned about AI through the media. Faculty were more likely to report not having a basic understanding of AI technologies (K2, $p$=.031). Students were more interested in AI in patient care training, while faculty were more interested in AI in teaching training (K2, $p$=0.001). Additionally, students and faculty reported comparable attitudes towards AI, limited AI literacy, and time constraints in the curriculum. There is interest in broad and deep AI topics. Findings in medical students and faculty parallel results from AI research in other professional groups published. The research conclusively proved interest among medical students and faculty in AI technology overall and its applications in health and medicine. Machine scores correlated more strongly with faculty ratings than faculty ratings did with each other (machine: .28–.53, faculty: .13–.33) and were less specific to the case. Machine scores and faculty ratings were equally correlated with medical knowledge, clinical cognition, and previous diagnostic rationale. Machine scores were more strongly associated with clinical communication than faculty ratings (.43 vs .31). |
| Kong et al. | 2021 | Singapore | 106 medical students | Cohort | To describe the supplementary use of VIP in the second-year class, where a pilot study was conducted. | Of the 106 participants, 87% agreed or strongly agreed that the use of the VIP helped them remember the content, while 69% of them felt that the VIP increased their confidence and competence in history-taking. The VIP was well-received by students and showed promise as a tool to complement history tutorial sessions, prior to students' encounters with SPs and real patients. Therefore, this trend demonstrated its potential as an alternative when clinical rotations were postponed or canceled. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Choi et al. | 2022 | South Korea | 30 medical students | Randomized controlled trial | To determine which of the 4 laryngoscopes, including the A-LRYNGO, a newly developed channel-type videolaryngoscope with an embedded artificial intelligence-based glottis guidance system, is suitable for novice medical students' tracheal intubation training using personal protective equipment (PPE). | The time to fully insert the endotracheal tube with the Macintosh laryngoscope did not change significantly ($p$=.177), but the remaining outcomes improved significantly after the practical workshop (all $p$<0.05). Despite being novice practitioners with no intubation experience and using PPE, channel-type videolaryngoscopes were associated with good intubation performance before the practical workshop (all $p$<0.001). The artificial intelligence-based glottis guidance system of the A-LRYNGO showed 93.1% accuracy, but 20.7% of trials were guided by vocal folds. |
| Fazllolahi et al. | 2022 | Canada | 70 medical students | Randomized controlled trial | To determine how virtual reality-based simulation (VOA) and remote expert instruction compare in the acquisition of skills, affective outcomes, and cognitive outcomes of students during surgical simulation training. | The ICEMS evaluated 350 practical resections, and both the ICEMS and OSATS evaluated 70 realistic resections. VOA significantly improved practical experience scores by 0.66 (95% CI, 0.55 to 0.77) points compared to the instructor group and by 0.65 (95% CI, 0.54 to 0.77) points compared to the control group ($p$<0.001). Realistic proficiency scores were significantly higher for the VOA group compared to the instructor (difference of 0.53 [95% CI, 0.40 to 0.67] points; $p$<0.001) and control (difference of 0.49 [95% CI, 0.34 to 0.61] points; $p$<0.001) groups. Average overall OSATS ratings were not statistically significant among the VOA (4.63 [95% CI, 4.06 to 5.20] points), instructor (4.40 [95% CI, 3.88 to 4.91] points), and control (3.86 [95% CI, 3.44 to 4.27] points) groups. However, in OSATS subscores, VOA significantly improved the average overall OSATS subscore compared to the control group (mean difference, 1.04 [95% CI, 0.13 to 1.96] points; $p$=0.02), while specialized instruction significantly improved OSATS subscores for instrument handling vs control (mean difference, 1.18 [95% CI, 0.22 to 2.14]; $p$=0.01). No significant differences were found in cognitive load, positive activation, and negative emotions. Therefore, VOA feedback demonstrated superior performance outcome and skill transfer, with equivalent OSATS ratings and cognitive and emotional responses compared to remote expert instruction, indicating advantages for its use in simulation training. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Agarwal, Sharma, Goswami. | 2023 | India | Nonexistent | Quantitative cross-sectional study | To evaluate and compare the applicability of ChatGPT, Bard, and Bing in generating reasoning-based MCQs for undergraduate MBBS (Bachelor of Medicine, Bachelor of Surgery) students on the subject of physiology. | ChatGPT and Bard generated 110 MCQs for the chosen competencies. However, Bing provided only 100 MCQs as it failed to generate them for two competencies. The validity of MCQs was rated as 3 (3-3) for ChatGPT, 3 (1.5-3) for Bard, and 3 (1.5-3) for Bing, showing a significant difference ($p < 0.001$) among the models. The difficulty of MCQs was rated as 1 (0-1) for ChatGPT, 1 (1-2) for Bard, and 1 (1-2) for Bing, with a significant difference ($p$=0.006). The reasoning ability required to answer the MCQs was rated as 1 (1-2) for ChatGPT, 1 (1-2) for Bard, and 1 (1-2) for Bing, with no significant difference ($p$=0.235). K was ≥ 0.8 for all three parameters in all three AI models. AI still needs to evolve to generate reasoning-based MCQs in medical physiology. ChatGPT, Bard, and Bing showed certain limitations. Bing generated significantly fewer valid MCQs, while ChatGPT generated significantly less difficult MCQs. |
| Al Kahf et al. | 2023 | France | 359 medical students | Randomized controlled trial | To assess the impact of Chatprogress, a chatbot-based game developed at the University Paris Descartes containing 8 pneumology cases, on students' success rates in their final exams. | Differences in scores on the pneumology subtest throughout the academic year were significantly greater between Players and Users vs. Controls (average score: 12.7/20 vs. 12.0/20, $p$=0.0104 and average score: 12.7/20 vs. 12.0/20, $p$=0.0365, respectively). This significant difference was also present in the overall scores of the PCC test: (average score: 12.5/20 vs. 12.1/20, $p$=0.0285 and 12.6/20 vs. 12.1/20, $p$=0.0355, respectively). Although no significant correlation was found between pneumology subtest scores and MS attendance parameters (number of games completed out of the 8 proposed to Users and number of times a User completed a game), there was a trend for a better correlation when users were assessed on a subject covered by Chatprogress. MS was also found to be fond of this teaching tool, asking for more pedagogical feedback even when they answered questions correctly. This randomized controlled trial is the first to demonstrate a significant improvement in student outcomes (both in the pneumology subtest and in the overall PCC exam) when they had access to Chatbots, and even more so when they actually used it. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Breeding et al. | 2023 | USA | 56 medical students | Qualitative cross-sectional study | To evaluate how medical students and laypeople assess the information generated by ChatGPT compared to an evidence-based resource on the diagnosis and management of 5 common surgical conditions. | Out of the 56 survey participants, 50.9% (n = 28) were medical students from the US and 49.1% (n = 27) were from the general population. Medical students reported that ChatGPT articles exhibited significantly more clarity (appendicitis: 4.39 vs 3.89, $p$=.020; diverticulitis: 4.54 vs 3.68, $p$<.001; SBO: 4.43 vs 3.79, $p$=.003; GI bleeding: 4.36 vs 3.93, $p$=.020) and better organization (diverticulitis: 4.36 vs 3.68, $p$=.021; SBO: 4.39 vs 3.82, $p$=.033) than evidence-based sources. However, for all 5 conditions, medical students found evidence-based passages to be more comprehensive than ChatGPT articles (cholecystitis: 4.04 vs 3.36, $p$=0.009; appendicitis: 4.07 vs 3.36, $p$=0.015; diverticulitis: 4.07 vs 3.36, $p$=0.015; small bowel obstruction: 4.11 vs 3.54, $p$=0.030; upper GI bleeding: 4.11 vs 3.29, $p$=0.003). Medical students perceived ChatGPT articles to be clearer and better organized than evidence-based sources on the pathogenesis, diagnosis, and management of 5 common surgical pathologies. However, evidence-based articles were rated as significantly more comprehensive. |
| Banerjee et al. | 2023 | India | Nonexistent | Quantitative cross-sectional study | To explore the effectiveness of ChatGPT in solving various reasoning-based questions across the five core concepts applied to different modules in the subject of physiology. | The overall average score for the modules (60 questions) was 3.72 ± 0.26, while the average score obtained for the core concepts (60 questions) was 3.68 ± 0.30. Additionally, statistically significant differences ($p$=0.05 for modules and $p$=0.024 for core concepts) were observed among various modules as well as core concepts. The significant differences observed in scores between various modules and core concepts highlight the variable performance of the same software tool. |
| Cold et al. | 2023 | Denmark | 20 medical students | Randomized controlled trial | To evaluate whether feedback from an AI-based automatic identification system of bronchial segments improves the end-of-training performance of novice bronchoscopists. | FG significantly outperformed in all three outcome measures (median difference, P value): diagnostic completeness (3.5 segments, $p$<0.001), structured progression (13.5 correct progressions, $p$<0.001), and procedure time (-214 seconds, $p$=0.002). Training guided by this new AI leads novices to perform more complete, systematic, and faster bronchoscopies. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Das et al. | 2023 | India | Nonexistent | Quantitative cross-sectional study | To analyze ChatGPT's ability to respond to first and second-order questions on the subject of microbiology. | The overall score of 96 responses was 4.04 ± 0.37 (median: 4.17, Q1-Q3: 3.88-4.34), with the average score of responses to first-order knowledge questions being 4.07 ± 0.32 (median: 4.17, Q1-Q3: 4-4.33) and that of responses to second-order knowledge questions being 3.99 ± 0.43 (median: 4, Q1-Q3: 3.67-4.33) (Mann-Whitney $p$=0.4). The score was significantly below the score of 5 (one-sample median test $p$<0.0001) and similar to 4 (one-sample median test $p$=0.09). Overall, there was variation in median scores obtained across eight topic categories in microbiology, indicating inconsistent performance across different topics. The study results indicate that ChatGPT is capable of answering both first-order and second-order knowledge-related questions in microbiology. The model achieved an accuracy of approximately 80%, and there was no difference in the model's ability to answer first-order questions compared to second-order knowledge questions. The findings of this study suggest that ChatGPT has the potential to be an effective tool for automated responses to questions in the field of microbiology. |
| Dhanvijay et al. | 2023 | India | Nonexistent | Quantitative cross-sectional study | To evaluate the performance of three LLMs, namely ChatGPT (3.5; free research version), Google Bard (Experiment), and Microsoft Bing (Precise), in responding to physiology vignette cases. | The overall scores for ChatGPT, Bing, and Bard in the study, with a total of 77 cases, were considered 3.19 ± 0.3, 2.15 ± 0.6, and 2.91 ± 0.5, respectively, p < 0.0001. Therefore, ChatGPT 3.5 (free version) achieved the highest score, Bing (Precise) had the lowest score, and Bard (Experiment) ranked between the two in terms of performance. The mean ICC values for ChatGPT, Bing, and Bard were 0.858 (95% CI: 0.777 to 0.91, p < 0.0001), 0.975 (95% CI: 0.961 to 0.984, p < 0.0001), and 0.964 (95% CI: 0.944 to 0.977, p < 0.0001), respectively. ChatGPT outperformed Bard and Bing in responding to case vignettes in physiology. Therefore, students and teachers may consider choosing LLMs for their educational purposes according to case-based learning in physiology. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Fazllolahi et al. | 2023 | Canada | 46 medical students e 14 specialists | Cohort | To evaluate the pedagogical value of technical skills selected by AI and their extended effects on surgical simulation training. | A total of 46 medical students (mean age [range], 22 [18-27] years; 27 [59%] women) and 14 surgeons (median age [range], 45 [35-59] years; 14 [100%] men) were included in this study, and no participants were lost to follow-up. Feedback on 4 technical skills selected by the AI was associated with additional performance changes in 32 metrics throughout the entire procedure and 20 metrics during tumor removal that were not observed in the control group. Participants exposed to the AI-enhanced curriculum demonstrated significant improvement in safety metrics, such as reduced healthy tissue removal rate (difference −7.05 × 10−5 [95% CI, −1.09 × 10−4 to −3.14 × 10−5] mm3 per 20 ms; $p$<0.001) and maintenance of focused bimanual control of the operative field (difference in maximum instrument divergence, −4.99 [95% CI, −8.48 to −1.49] mm, $p$=0.006) compared with the control group. However, unintended negative effects were also observed. These included significantly slower speed and acceleration in the dominant hand (speed: mean difference, −0.13 [95% CI, −0.17 to −0.09] mm per 20 ms; $p$<0.09]; acceleration: mean difference, −2.25 × 10−2 [95% CI, −3.20 × 10−2 to −1.31 × 10−2] mm per 20 ms2; $p$<0.001) and a significant reduction in tumor removal rate (mean difference, −4.85 × 10−5 [95% CI, −7.22 × 10−5 to −2.48 × 10−5] mm3 per 20 ms; $p$<0.001) compared with the control. These unintended results diverged students' movement performance and efficiency metrics from specialization benchmarks. In this study, an AI-enhanced curriculum for bimanual surgical skills resulted in unintended changes that improved safety performance but adversely affected some efficiency metrics. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Friederichs, Friedrichs, März. | 2023 | Germany | Nonexistent | Quantitative cross-sectional study | To measure the level of factual knowledge in ChatGPT responses, comparing ChatGPT's performance with that of medical students on a progress test. | Out of the 395 evaluated responses, 65.5% of the progress test questions answered by ChatGPT were correct. On average, ChatGPT required 22.8 seconds (SD 17.5) for a complete response, containing 36.2 (SD 28.1) words. There was no correlation between the time taken and word count with the accuracy of ChatGPT responses (correlation coefficient for time rho = −0.08, 95% CI [−0.18, 0.02], t(393) = −1.55, $p$=0.121; for word count rho = −0.03, 95% CI [−0.13, 0.07], t(393) = −0.54, $p$=0.592). There was a significant correlation between the difficulty index of the MCQs and the accuracy of ChatGPT responses (correlation coefficient for difficulty: rho = 0.16, 95% CI [0.06, 0.25], t(393) = 3.19, $p$=0.002). ChatGPT was able to correctly answer two-thirds of all MCQs at the level of the state licensing exam in German Progress Test Medicine and outperformed nearly all medical students in years 1 to 3. |
| Ghosh, Bir. | 2023 | India | Nonexistent | Quantitative cross-sectional study | To determine if ChatGPT can solve higher-order problems related to medical biochemistry. | The AI software responded to 200 questions that required higher-order thinking with a median score of 4.0 (Q1=3.50, Q3=4.50). Using a single-sample Wilcoxon signed-rank test, the result was lower than the hypothetical maximum of five ($p$=0.001) and comparable to four ($p$=0.16). There was no difference in responses to questions from different CBME modules in medical biochemistry (Kruskal-Wallis $p$=0.39). The reliability between classifiers two professors. The results of this research indicate that ChatGPT has the potential to be a successful tool for answering questions that require higher-order thinking in medical biochemistry, with a median score of four out of five. |
| Hollis-Sando et al. | 2023 | Australia | 181 medical students | Mixed methods study | To assess both student opinion regarding the acceptability of computer-based scoring (machine learning), and a preliminary study of the accuracy of NLP implemented with open-source software in scoring SAQs of medical students. | Although students expressed relatively concern about the accuracy of computer-based marking, the majority of students agreed that computer marking would be more objective than human marking (67.0%) and reported that they would not oppose computer-based marking (55.5%). Regarding automated marking of SAQs, for 1-point questions, there were consistently high classification accuracies (average precision of 0.98). For more complex 2 and 3-point SAQs, where multiclass classification was required, the accuracy was lower (average 0.65 and 0.59, respectively). Medical students may support computer-based marking due to its objectivity. DL has the potential to provide accurate marking of written questions. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Huh S | 2023 | South Korea | 77 medical students | Quantitative cross-sectional study | To compare the knowledge and interpretative ability of ChatGPT, a general artificial intelligence language model, with those of medical students in Korea by administering a parasitology exam to ChatGPT and medical students. | The ChatGPT answered correctly 48 out of 79 items (60.8%). This score was lower than the average score of 77 medical students, which was 71.8 out of 79 (90.8%), with a minimum score of 65 (89.0%) and a maximum score of 74 (93.7%). The ChatGPT's performance was lower than that of the medical students. The correct response rate shown by ChatGPT was not related to the difficulty level of the items. However, there was an association between acceptable explanations and correct responses. The ChatGPT's correct response rate of 60.8% was not necessarily an indicator of poor performance, as the questions were not easy for medical students to answer correctly. The considerably higher average score (89.6%) of medical students may have been due to their prior learning in parasitology and the fact that the exam was administered 4 days after the class. The knowledge and interpretive ability of ChatGPT in responding to this parasitology exam are not yet comparable to those of medical students in Korea. |
| Ignjatovic', Stevanovic'. | 2023 | Serbia | Nonexistent | Quantitative cross-sectional study | To evaluate the performance of ChatGPT (GPT-3.5 and GPT-4) as a study tool in solving biostatistical problems and identify any potential disadvantages that may arise from the use of ChatGPT in medical education, particularly in practical biostatistical problem-solving. | GPT-3.5 successfully solved 5 practical problems on the first attempt, related to categorical data, cross-sectional study, measurement reliability, probability properties, and the t-test. However, it did not provide correct answers regarding analysis of variance, chi-square test, and sample size within 3 attempts. GPT-4 also solved a task related to confidence interval on the first attempt and answered all questions within 3 attempts, with accurate guidance and monitoring. Assessment of both versions' performance on 10 bio-statistical problems revealed that the performance of GPT-3.5 and 4 was below average, with correct response rates of 5 and 6 out of 10 on the first attempt. GPT-4 managed to provide all correct answers within 3 attempts. These findings indicate that students should be aware that this tool, while capable of providing and calculating different statistical analyses, may be incorrect, and they should be mindful of the limitations of ChatGPT and exercise caution when incorporating this model into medical education. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Ilgaz, Çelik. | 2023 | Türkiye | Nonexistent | Quantitative cross-sectional study | To assess how various open AI platform resources, such as ChatGPT and Google Bard, in their current form, can contribute to anatomy education with a perspective of questioning, answering, and writing articles, aiming to address some controversial topics. | The most successful results were observed in the first user of ChatGPT in Turkish (48.09%), followed by the second user of Google Bard in Turkish (45.80%), and the others were the first user in Google Bard in Turkish (45.03%), ChatGPT in English (44.27%), Google Bard in English (41.98%), and the first user of ChatGPT in Turkish (41.22%), respectively. None of them achieved a score above 50% in the total of correct answers. When the overall results were compared across all groups, no statistically significant difference was found ($p>0.05$). There were differences in the performance of the LLMs, but no significant difference was observed between ChatGPT and Google Bard in terms of question answering. The issue that generates these models was addressed. It may be useful for exam preparation in educational aspects. The performance of the applications in article writing was not yet at a sufficient level. Although LLMs show promise as valuable tools for medical education, their use requires caution, given the complex algorithms and their limitations in dealing with potential inaccuracies. |
| Meo et al. | 2023 | Saudi Arabia | Nonexistent | Quantitative cross-sectional study | To investigate the level of knowledge of ChatGPT in medical education, both in basic medical sciences and clinical sciences, in exam-based performance of multiple-choice questions (MCQs) and its impact on the medical examination system. | The results revealed that out of 100 MCQs in various disciplines of basic and clinical medical sciences, ChatGPT attempted all MCQs and scored 37/50 (74%) in basic medical sciences and 35/50 (70%) in clinical medical sciences, with an overall score of 72/100 (72%) in basic and clinical medical sciences. It is concluded that ChatGPT achieved a satisfactory score in both basic and clinical medical science disciplines and demonstrated a degree of understanding and explanation. The findings of this study suggest that ChatGPT may be able to assist medical students and educators in medical education settings, as it holds potential as an innovation in the field of medical science and education. |

continue...

... continuation Appendix 1

| Author | Year | Country | Size Sample | Study Type | Objective | Outcome |
|---|---|---|---|---|---|---|
| Morjaria et al. | 2023 | Canada | 10 medical students | Quantitative cross-sectional study | To evaluate the extent of threat posed by ChatGPT to the validity of short-answer assessment problems used to examine pre-clinical medical students in an undergraduate medical education program. | The responses generated by ChatGPT received an average score of 3.29 out of 5 (n = 30, 95% CI 2.93-3.65) compared to 2.38 for a group of students who achieved minimum passing grades (n = 10, 95% CI 1.94-2.82), representing higher performance ($p$=0.008, κ2 = 0.169), but were surpassed by the historical class average scores on the same 30 problems (mean 3.67, $p$=0.018) when including all past responses, regardless of the student's performance level. There was no statistically significant trend in performance across all domains of Bloom's Taxonomy. Although ChatGPT was able to pass short-answer assessment problems covering the pre-creation curriculum, it only outperformed low-performing students. We observed that tutors in various cases were convinced that the responses produced by ChatGPT were generated by the students. Risks to the validity of the assessment include uncertainty in identifying students with difficulties and the inability to intervene in a timely manner. |
| Roos et al. | 2023 | Germany | Nonexistent | Quantitative cross-sectional study | To assess and compare the performance of 3 LLMs, GPT-4, Bing, and GPT-3.5-Turbo, in the 2022 German State Medical Exams and assess their performance in comparison to that of medical students. | GPT-4 achieved the highest overall performance, correctly answering 88.1% of the questions, closely followed by Bing (86.0%) and GPT-3.5-Turbo (65.7%). Students had an average correct response rate of 74.6%. Both GPT-4 and Bing significantly outperformed students in both exams. When media-related questions were excluded, Bing achieved the highest performance at 90.7%, closely followed by GPT-4 (90.4%), while GPT-3.5-Turbo lagged behind (68.2%). There was a significant decline in the performance of GPT-4 and Bing in the fall 2022 exam, attributed to a higher proportion of media-related questions and a potential increase in question difficulty. LLMs, particularly GPT-4 and Bing, demonstrate potential as valuable tools in medical education and for pre-test exam questions. |

**Source:** Elaborated by authors.