revista docência do ensino superior

VOLUME 15
2025
ISSN: 2237-5864

DOI: https://doi.org/10.35699/2237-5864.2025.58339

SEÇÃO: IA NOS PROCESSOS DE ENSINO-APRENDIZAGEM

How generative artificial intelligence can facilitate the teaching of clinical reasoning: a scoping review

Como a inteligência artificial generativa pode facilitar o ensino do raciocínio clínico: uma revisão de escopo

Cómo la inteligencia artificial generativa puede facilitar la enseñanza del razonamiento clínico: una revisión exhaustiva

Guilherme Freitas Bernardo Ferreira[1], Alexandre Sampaio Moura[2], Lígia Maria Cayres Ribeiro[3], Maria Aparecida Turci[4], Sílvia Mamede[5]

ABSTRACT

This review aims to map and summarize the current state of research to identify the applicability of chatbots in teaching clinical reasoning during medical training, considering the best available evidence. A systematic and comprehensive search was conducted in PubMed/MEDLINE, Web of Science, and Google Scholar databases between August 2023 and August 2024. Original studies describing educational applications aligned with evidence-based strategies for teaching clinical reasoning (self-explanation, structured reflection, case practice, and feedback) were included. The selection was complemented by snowballing and expert consultation. Twenty-one publications were included. All studies explored the use of ChatGPT (OpenAI); three (14%) also analyzed Bard (Google), two (9.5%) investigated Bing (Microsoft),

[1] Universidade Professor Edson Antonio Velano (Unifenas-BH), Belo Horizonte, MG, Brasil.
ORCID ID: https://orcid.org/0009-0006-2383-7462. E-mail: gfreitasbernardo@gmail.com
[2] Faculdade Santa Casa BH, Belo Horizonte, MG, Brasil.
ORCID ID: https://orcid.org/0000-0002-4818-5425. E-mail: alexandremoura@faculdadesantacasabh.edu.br
[3] University Medical Center Groningen, Groningen, Netherlands.
ORCID ID: https://orcid.org/0000-0002-4277-3066. E-mail: l.m.cayres.ribeiro@umcg.nl
[4] Universidade Professor Edson Antonio Velano (Unifenas-BH), Belo Horizonte, MG, Brasil.
ORCID ID: https://orcid.org/0000-0002-4380-4231. E-mail: mariaturci@gmail.com
[5] University Medical Center Groningen, Groningen, Netherlands.
ORCID ID: https://orcid.org/0000-0003-1187-2392. E-mail: s.mamede@erasmusmc.nl

Recebido em: 23/04/2025 *Aprovado em: 17/11/2025* *Publicado em: 30/12/2025*

and one (5%) explored other artificial intelligence tools. Our findings suggest that chatbots can support the development of clinical reasoning skills through effective educational strategies. Chatbot responses can help students build understanding, promote deliberate reflection, encourage feedback when practicing with written cases, and adapt content to the learner's stage. Few studies raised concerns about risks and ethical issues. This review demonstrated that chatbots hold great potential to enhance the development of clinical reasoning during medical education. However, it is essential to address inherent limitations, such as the risks of hallucinations and inaccurate explanations, to maximize the technology's educational potential.

Keywords: clinical reasoning; health education; artificial intelligence; generative artificial intelligence; scoping review.

RESUMO

Esta revisão tem como objetivo mapear e resumir o estado atual da pesquisa para identificar a aplicabilidade dos chatbots no ensino do raciocínio clínico durante a formação médica, considerando as melhores evidências disponíveis. Foi realizada uma busca sistemática e abrangente nas bases de dados PubMed/MEDLINE, Web of Science e Google Scholar, entre agosto de 2023 e agosto de 2024. Foram incluídos estudos originais que descreveram aplicações educacionais alinhadas a estratégias com evidência para o ensino do raciocínio clínico (autoexplicação, reflexão estruturada, prática com casos e feedback). A seleção foi complementada por *snowballing* e consulta a especialistas. Foram incluídas 21 publicações. Todos os estudos exploraram o uso do ChatGPT (OpenAI); três (14%) também analisaram o Bard (Google), dois (9,5%) investigaram o Bing (Microsoft) e um (5%) explorou outras ferramentas de inteligência artificial. Nossos achados sugerem que chatbots podem apoiar o desenvolvimento de habilidades de raciocínio clínico por meio de estratégias educacionais eficazes. As respostas dos chatbots podem ajudar os estudantes a construir compreensão, promover reflexão deliberada, incentivar feedback ao praticar com casos escritos e adaptar o conteúdo ao estágio de aprendizagem. Poucos estudos levantaram preocupações sobre riscos e questões éticas. Esta revisão demonstrou que os chatbots apresentam um grande potencial para aprimorar o desenvolvimento do raciocínio clínico durante a formação médica. No entanto, é fundamental abordar as limitações inerentes, como os riscos de alucinações e explicações imprecisas, para maximizar o potencial educacional da tecnologia.

Palavras-chave: raciocínio clínico; ensino em saúde; inteligência artificial; inteligência artificial generativa; revisão de escopo.

RESUMEN

Esta revisión tiene como objetivo mapear y resumir el estado actual de la investigación para identificar la aplicabilidad de los chatbots en la enseñanza del razonamiento clínico durante la formación médica, considerando las mejores evidencias disponibles. Se realizó una búsqueda sistemática y exhaustiva en las bases de datos PubMed/MEDLINE, Web of Science y Google Scholar entre agosto de 2023 y agosto de 2024. Se incluyeron estudios originales que describieran aplicaciones educativas alineadas con estrategias basadas en evidencia para la enseñanza del razonamiento clínico (autoexplicación, reflexión estructurada, práctica con casos y retroalimentación). La selección se complementó mediante la técnica de snowballing y consulta a expertos. Se incluyeron 21 publicaciones. Todos los estudios exploraron el uso de ChatGPT (OpenAI); tres (14%) también analizaron Bard (Google), dos (9,5%) investigaron Bing (Microsoft) y uno (5%) exploró otras herramientas de inteligência artificial. Nuestros hallazgos sugieren que los chatbots pueden apoyar el desarrollo de habilidades de razonamiento clínico mediante estrategias educativas efectivas. Las respuestas de los chatbots pueden ayudar a los estudiantes a construir comprensión, promover la reflexión deliberada, fomentar la retroalimentación al practicar con casos escritos y adaptar el contenido al nivel de aprendizaje. Pocos estudios abordaron preocupaciones relacionadas con riesgos y cuestiones éticas. Esta revisión demostró que los chatbots presentan un gran potencial para mejorar el desarrollo del razonamiento clínico durante la formación médica. No obstante, es fundamental abordar las limitaciones inherentes, como los riesgos de alucinaciones y explicaciones imprecisas, para maximizar el potencial educativo de la tecnología.

Palabras clave: razonamiento clínico; enseñanza en salud; inteligencia artificial; inteligencia artificial generativa; revisión de alcance.

INTRODUCTION

Since the release of generative artificial intelligence chatbots (GAIC) such as ChatGPT (OpenAI), many attempts have emerged to explore their use in patient care and medical education. The potential use of GAIC as supporting tools for physicians has been extensively explored. A similar trend has been observed in medical training (Gordon *et al.*, 2024). While the responsible use of GAIC in clinical practice depends critically on human clinical reasoning expertise, developing such expertise can also be a benefit from this technology. Nevertheless, whether and how GAIC can help teach clinical reasoning during medical training remains unclear.

Generative artificial intelligence refers to computational systems capable of producing text, images, or code, based on large language models (LLMs) trained on extensive datasets. Chatbots using this technology, referred to as GAIC, simulate natural human conversation,

interpret context, and generate coherent responses. The most prominent examples include ChatGPT (OpenAI, 2022), Bard (Google, 2023, later integrated into Gemini), and Bing Chat (Microsoft, 2023).

Clinical reasoning is a complex set of skills, processes, or outcomes wherein clinicians observe, collect and interpret data to diagnose and treat patients (Mamede, 2020). Therefore, it is essential in a doctor's performance and crucial for diagnostic and therapeutic accuracy.

Previous studies have shown that expertise in medicine develops through a process where biomedical knowledge becomes integrated with a clinical one into illness scripts (Schmidt; Rikers, 2007). These 'packages' are cognitive schemas that become increasingly refined through experience with clinical problems. This is so because clinical reasoning relies on an extensive base of acquired knowledge and depends on constructing and activating relevant mental *scripts* (Bowen, 2006; Cutrer; Sullivan; Fleming, 2013; Eva, 2005; Mamede, 2020).

With that in mind, teachers should be aware that it is impossible to transfer the ability to reason to solve problems linearly. Instead, a more comprehensive approach to clinical teaching should be taken, focusing on fostering the development of large sets of illness scripts (Eva, 2005; Mamede, 2020). Effective teaching strategies should emphasize the importance of using diverse examples to build a robust mental database for students, integrate biomedical and clinical concepts, mimic real-life scenarios, and improve understanding and diagnosis (Eva, 2005).

Few strategies have been proven effective to teach clinical reasoning (Prakash; Sladek; Schuwirth, 2019). Some of the interventions with proven benefits are strategies that (1) build an understanding of causal mechanisms of diseases, such as *self-explanation,* (2) foster comparing and contrasting alternative diagnoses such as structured or deliberated reflection, (3) enable practice with entire cases with provision of feedback, (4) employ retrieval practice, (5) promote learning by comparing and contrasting discriminating features of different diagnosis, and (6) match the student's stage of learning (Cooper *et al.*, 2021).

The educational challenge of teaching clinical reasoning can benefit from new technologies. In particular, the recent emergence of GAIC sparked increased interest in their applications in medical education (Lee, 2023). A scoping review conducted by Preiksaitis and Rose (2023) revealed that artificial intelligence (AI) holds transformative potential for medical education as it offers exciting opportunities. The authors have identified that AI could help improve understanding, work as a continuous education and self-directed learning tool, develop personalized learning plans, and provide feedback (Preiksaitis; Rose, 2023).

GAIC may, therefore, support the development of clinical reasoning if they can facilitate one or more strategies that have proven successful in this task. This review mapped, reviewed, and summarized the state of current research to identify the applicability of GAIC for teaching clinical reasoning during medical training based on the best evidence on how to facilitate it.

METHODS

Overview

A scoping review of the literature followed the framework proposed by Arksey and O'Malley (2007).

Research question

The primary research question to guide the review was: "How can GAIC foster the development of clinical reasoning during medical training?" The concept of the potential use of GAIC in educational strategies adopted in this review was to help medical students develop clinical reasoning for their future performance. This means that our focus was on using GAIC as clinical reasoning-supporting tools in education, not in clinical practice.

Search for research evidence

We searched for the articles in three databases: World of Science (WoS), PubMed and Google Scholar. The search strategy and keywords are displayed in Appendix 1. We restricted our search to original articles in English to ensure terminological consistency and methodological comparability across sources. Most of the recent research on generative AI chatbots in medical education has been published in English.

Bibliographies of the studies found through database searches were screened to identify further references. We also used existing knowledge and networks of experts to obtain titles that could meet the selection criteria adopted.

Team and roles

Authors 1 (G.F.B.F.) and 2 (A.S.M.) searched for and selected the included articles, extracted the primary information contained in the selected articles, and wrote the manuscript. Authors 3 (L.M.C.R.) and 5 (S.M.) guided the theoretical foundation for the research and reviewed the paper. Author 4 (M.A.T.) designed the search and selection method and reviewed the draft of the manuscript. All authors participated in the analysis and interpretation of the results,

discussed the relevance of the data, and established the relationship between GAIC capabilities and their educational potential in the teaching of clinical reasoning.

Study selection

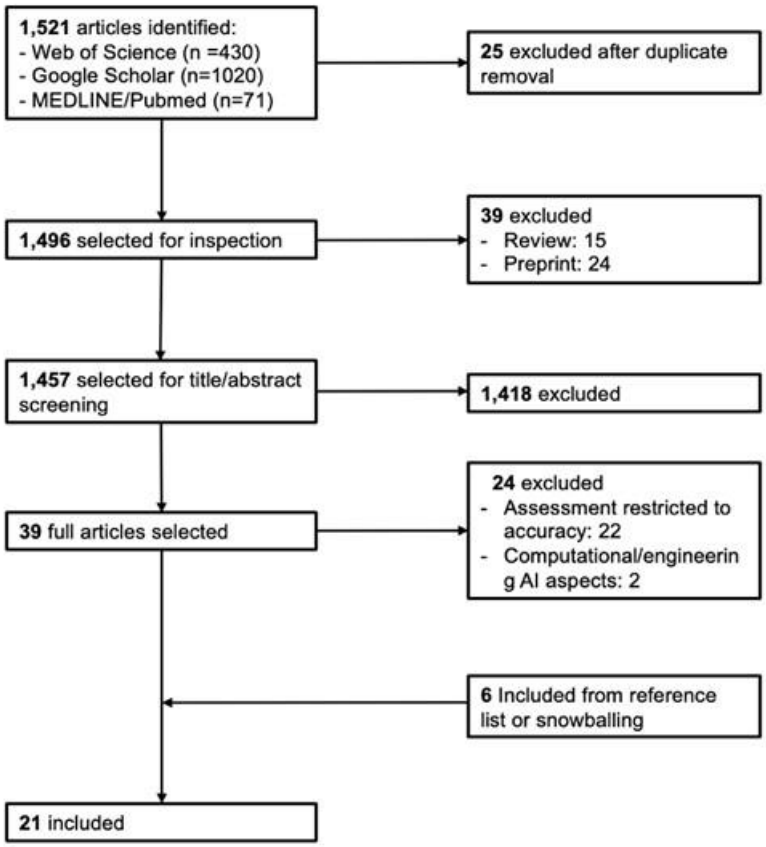
We selected original research articles with primary or secondary outcomes that could be useful in designing or implementing educational activities based on strategies that have proved effective for teaching clinical reasoning. As "effective", we considered educational strategies that have been pointed out by reviews (Cooper *et al.*, 2021; Prakash; Sladek; Schuwirth, 2019) as such: 1) promotes structured reflection, 2) builds understanding through explanations on the reasoning behind solving a clinical vignette/self-explanation, 3) promotes comparing/contrasting between differential diagnosis, 4) encourages practice with cases with the provision of feedback. We also included articles that investigated tools that could help reduce teachers' workload. We excluded preprints, perspective articles, publications on computational aspects of AI, and studies on diagnostic accuracy not linked to the teaching-learning environment.

Two independent researchers, authors 1 (G.F.B.F.) and 2 (A.S.M.), initially selected the articles. Those were screened according to title and abstract for relevance and were selected after a full read. In addition to the database searches, two experts in medical education, authors 3 (L.M.C.R.) and 5 (S.M.) were consulted via email. Each expert suggested additional studies consistent with the predefined selection criteria. Six of these publications were included in the final sample after reviewers' consensus.

RESULTS

Our search initially identified 1,521 publications. After the exclusion of duplicates, preprints, and perspective articles, we screened the remaining 1,457 ones according to title and abstracts. We then excluded 1,418 papers that were considered unrelated to our research. The remaining 39 articles were selected for full reading, and 24 ones were excluded. Other six articles were included by snowballing or searching the list of references after the concordance of the two authors. Thus, 21 texts were included in this review. The search and selection process are summarized in Figure 1.

Figure 1 – Flow Diagram



Source: made by the authors, 2025.

Studies characteristics

All studies explored ChatGPT (OpenAI); three (14%) additionally explored the use of Bard (Google), two (9,5%), the use of Bing (Microsoft); and, one (5%) explored other AI tools. All studies were published in 2023 and 2024.

Five (24%) articles reported chatbot diagnostic accuracy and differential diagnostic lists generation (Balas; Ing, 2023; Hirosawa *et al.*, 2023; Kanjee; Crowe; Rodman, 2023; Koga; Martin; Dickson, 2023; Shea *et al.*, 2023), ten (48%) focused on the performance in standard tests or exams (Bonetti *et al.*, 2023; Balasanjeevi; Surapaneni, 2024; Cai *et al.*, 2023; Fonseca *et al.*, 2024; D'Souza *et al.*, 2023; Hirosawa *et al.*, 2023; Kung *et al.*, 2023; Madrid-García *et al.*, 2023; Shieh *et al.*, 2024; Yiu; Lam, 2023), one (5%) measured clinical reasoning skills as the primary outcome (Strong *et al.*, 2023), two (14%) analyzed chatbot performance on simulation scenarios (one as a tutor and one as a participant) (Scherr *et al.*, 2023; Xie *et al.*, 2023), and four (19%) evaluated its ability to write clinical exams (KLANG *et al.*, 2023; Hudon *et al.*, 2024; Kiyak; Emekli, 2024; Wong *et al.*, 2024). Relevant data were charted, summarized, and reported in Table 1 (placed after Conclusion).

Diagnostic accuracy and differential diagnostic lists

GAIC showed overall good diagnostic accuracy, at least comparable to experienced physicians in Internal Medicine, Ophthalmology, and Geriatrics. GPT-4.0 performed consistently better than its previous versions and was overall more accurate than other chatbots (Koga; Martin; Dickson, 2023).

Differential diagnostic lists elaborated by GAIC were usually comprehensive and included up to ten items. However, this number varied according to the prompt used. These lists included a high rate of correct diagnoses and were considered accurate and coherent with the presented clinical cases. Of the GAIC investigated, GPT-4 was considered more accurate than GPT-3.5 and Google Bard (Koga; Martin; Dickson, 2023). Three (60%) articles reported risks of bias and hallucinations (i.e., when GAIC generate factually incorrect answers) when interacting with chatbots (Balas; Ing, 2023; Hirosawa *et al.*, 2023; Kanjee; Crowe; Rodman, 2023).

Test performance

Ten (48%) studies investigated how GAIC would perform in standard tests in different contexts. Three (50%) reported that GAIC achieved passing rates in board certification exams, including USMLE (Cai *et al.*, 2023; Kung *et al.*, 2023; Yiu; Lam, 2023). One has found that GPT-3 would be admitted to any residency program in Italy (Bonetti *et al.*, 2023).

The tests varied significantly regarding the types of questions (e.g., multiple-choice vs. open-ended), reasoning levels (single-step vs. multiple-steps), and fields of knowledge. GPT-4.0 performed better than its previous versions and other GAIC.

Noteworthy, GAIC usually provided explanations for the questions solved. Investigators scrutinized these explanations and often considered them coherent and appropriate (Bonetti *et al.*, 2023; Balasanjeevi; Surapaneni, 2024; Fonseca *et al.*, 2024; Kung *et al.*, 2023). However, some concerns were raised, such as those presented by Cai *et al.* (2023), where ChatGPT provided “an accurate description of trabeculectomy” but neglected “to mention that an aqueous shunt is the preferred procedure in this specific scenario, which could impact clinical decision-making” (Cai *et al.*, 2023, p. 145). In their study, hallucination was also frequent (18%). Yiu and Lam (2023, p. 6) noted that “LLMs achieve high concordance and provide insightful responses to test questions,” [but]“inappropriate or inaccurate decision-making, incomplete appreciation of nuanced clinical scenarios and utilization of out-of-date guidance was, however, noted”.

Clinical reasoning skills

Strong *et al.* (2023) compared ChatGPT's performance with medical students in a clinical reasoning test. The authors reported that GPT-4.0 scored significantly better than GPT-3.5 and was similar to graduate medical students. Passing rates were similar between GPT-4.0 and students (93% vs 85%). Both were significantly higher than GPT-3.5 (43%). The authors showed that GPT-4.0 could provide better problem lists, and other skills (diagnostic schema, differential diagnosis, illness scripts, case summary) matched those presented by the students in this test.

Simulation scenarios

Two articles evaluated GAIC abilities in designing simulated clinical cases. One (Scherr *et al.*, 2023) investigated whether ChatGPT could act as a tutor to provide instructions in a written case simulation, give feedback, and adapt case progression according to the user's responses. Although it suggests it can be a promising tool in the simulation field, there were some limitations regarding inaccuracy, technical errors, and risk of bias. The authors mentioned that "ChatGPT occasionally provided feedback on the appropriateness of a decision and then progressed the simulation as if the correct decision had been made" (Scherr *et al.*, 2023, p. 9) and considered some of the feedback weak and unclear.

Another study (Xie *et al.*, 2023) explored ChatGPT, Bard, and Bing's performances as respondents to increasingly complex written clinical scenarios. The authors showed that of the three models, ChatGPT is currently the most reliable regarding comprehensibility and alignment with clinical guidelines. They also alerted for "misleading or inaccurate responses due to biases in the training data or misconceptions of intricate medical concepts" (Xie *et al.*, 2023, p. 9).

Writing of clinical exams

Four studies explored the potential of GPT-4.0 to create medical examinations (KLANG *et al.*, 2023; Hudon *et al.*, 2024; Kiyak; Emekli, 2024; Wong *et al.*, 2024). After being tutored with an existing model, the machine made a 210 multiple-choice medical examination. Only 0,5% of the questions were labeled entirely inaccurate by the investigators, and 15% required some revision. Two authors explored ChatGPT`s ability to craft Scripts Concordance Tests and demonstrated that although promising (33), severe limitations existed, such as caricatural or stereotypical clinical presentations (Hudon *et al.*, 2024). Hudon *et al.* (2024) also noted that users could not accurately tell if a human or GAIC created a test.

DISCUSSION

This review aimed to better understand how GAIC could help teach clinical reasoning during medical training, considering the best evidence available on clinical reasoning development. Our findings suggest that GAIC have overall good diagnostic accuracy when dealing with written clinical vignettes, provide coherent differential diagnostic lists (Balas; Ing, 2023; Hirosawa *et al.*, 2023; Kanjee; Crowe; Rodman, 2023; Koga; Martin; Dickson, 2023; Shea *et al.*, 2023; Shieh *et al.*, 2024) alongside with explanations to their answers in clinical tests (Bonetti *et al.*, 2023; Balasanjeevi; Surapaneni, 2024; Cai *et al.*, 2023; Fonseca *et al.*, 2024; D'Souza *et al.*, 2023; Kung *et al.*, 2023; Madrid-García *et al.*, 2023; Shieh *et al.*, 2024; Strong *et al.*, 2023; Yiu; Lam, 2023), match clinical reasoning skills similar to those students that received specific training (Strong *et al.*, 2023), and can interact as a tutor in previously trained simulation scenarios (Scherr *et al.*, 2023; Xie *et al.*, 2023). Based on findings of recent systematic reviews (Cooper *et al.*, 2021; Prakash; Sladek; Schuwirth, 2019) that pointed out evidenced-based strategies to teach clinical reasoning, we aimed to correlate these strategies with the skills demonstrated by GAIC.

Strategies aimed at building understanding

The explanations provided by GAIC could be confronted with a self-explanation given by the student during practice with clinical cases (Cooper *et al.*, 2021). Interestingly, GAIC provided reasons that contained nonobvious insights, requiring deduction or external knowledge to the question input (Kung *et al.*, 2023). Also, Strong *et al.* (2023) investigated ChatGPT's clinical reasoning ability and demonstrated that it is at least comparable to that of students with formal training. The ability shown by GAIC to summarize clinical data and describe the reasoning behind the provided responses can act as a source of feedback, helping students reflect on and refine their explanations for a problem. This may contribute to restructuring knowledge of causal mechanisms underlying diseases and refining illness scripts according to their level of expertise. It has been demonstrated that knowledge of causal mechanisms acts as a "glue" that helps link clinical findings together, thereby facilitating the recognition of diseases (Woods *et al.*, 2006). This can ultimately increase diagnostic accuracy and provide better results for patients in the future.

However, problems inherent to AI, such as hallucinations and providing inaccurate explanations, may limit its use, at least in the models currently available (Balas; Ing, 2023; Hirosawa *et al.*, 2023; Kanjee; Crowe; Rodman, 2023).

Promote reflection

Structured reflection during practice with clinical cases has proven to improve students' diagnostic performance, possibly through refinements of illness scripts and diagnostic schemas (Cooper *et al.*, 2021; Mamede; Schmidt, 2023; Prakash; Sladek; Schuwirth, 2019). By generating a broad, accurate differential diagnosis list, GAIC can potentially help guide the students' reflection. These lists not only can help students identify which diseases they should consider when reflecting upon a particular clinical case but also help focus their study on discriminating/defining factors, similar to what happened in researches that used a "cued" reflection (Ibiapina *et al.*, 2014). We have found that the lists provided by GAIC are sufficiently accurate and reliable. Still, they should only be used when guided by specialists or teachers due to the risk of bias and issues mentioned above.

Practice with entire cases with the provision of feedback

Practice with a large sample of clinical problems is considered critical for developing clinical reasoning (Eva, 2005). Moreover, research in many domains has demonstrated that practice with problems associated with corrective feedback is the primary mechanism for developing expertise (Ericsson, 2004). It is likely, therefore, that providing students with feedback when they practice with clinical cases would be beneficial. Scherr *et al.* (2023) demonstrated that GPT-3.5 could act as a conductor in two critical care common scenarios with post-scenario feedback. Although the benefits of simulation-based learning are beyond the scope of this review, we believe that by doing so, GAIC can help with self-directed learning. We highlight that the risks of hallucinations and the limited number of studies exploring this type of interaction with GAIC impose significant limitations.

Adaptation to the stage of learning

Teaching should be tailored appropriately to each stage of learning (Cooper *et al.*, 2021). Koga, Martin and Dickson (2023) suggest that GAIC can facilitate discussions by novice participants when participating in complex clinicopathologic discussions. Kung *et al.* (2023, p. 9) argued that GAIC could provide "nonobvious concepts that may not be in learners' sphere of awareness". Also, GAIC adaptive tutoring was discussed above. We believe these characteristics give GAIC the potential to help foster active learning and consolidate illness scripts and mental schemas.

Concerns regarding AI and clinical reasoning education

A few studies mentioned concerns about ethical aspects, risks for students and patients, and the exacerbation of inequalities related to access to available technologies (D'Souza *et al.*,

2023; Kung *et al.*, 2023; Xie *et al.*, 2023; Yiu; Lam, 2023). One study noted that due to the nature of GAIC, meaning that they extract information from web-based sources, their generated clinical scenarios and responses "may reflect societal and systemic biases that already exist in medical education" (Scherr *et al.*, 2023, p. 10). In addition, some studies call attention to privacy and data security threats, especially when considering learning in real clinical settings (Civaner *et al.*, 2022; Grunhut; Wyatt; Marques, 2021; Lee, 2023). Lastly, most published studies on AI and clinical reasoning aimed to report the machine's capabilities to pass a standard test or establish correct diagnoses. Only a few studies sought to delve into the interaction between students and artificial intelligence to understand how these can enhance clinical reasoning skills among students. This would be critical to guide the incorporation of the tools in education. Nevertheless, it seems clear, particularly considering the limitations of the present AI models, that teachers should use technology as an ally in clinical reasoning education, not as a substitute.

Limitations?

Our review was limited by a lack of homogeneity in the terms referring to GAIC in the literature databases (e.g., large language models, generative artificial intelligence, chatbots, among others). This might have reduced the accuracy of the search and can explain the large number of excluded articles from the initial search string and the inclusion of a proportionally large number of manuscripts from reference lists and snowballing. It is worth noting that the word "chatbot" will be included as a Medical Subject Headings (MeSH) term in 2025.

CONCLUSION

This review has demonstrated that GAIC hold significant promise for enhancing clinical reasoning during medical training. They exhibit good diagnostic accuracy, generate coherent differential diagnoses, and provide detailed explanations to help students reflect and acquire knowledge. These capabilities make chatbots potentially powerful partners in implementing evidence-based strategies for teaching clinical reasoning. However, addressing inherent limitations, such as the risks of hallucinations and inaccurate explanations, is crucial to maximizing their educational potential.

Table 1 – Summary of findings

Authors	Country	Year published	Research method	Outcomes/ Questions asked	Chatbot	Main results	Relevance for clinical reasoning
Balas and Ing	Canada	2023	Quantitative	Providing the most likely and the differential diagnoses for ophthalmologic clinical vignettes	GPT-3	GPT-3 had a 90% diagnostic accuracy. The correct diagnosis was included in all differential diagnosis lists generated and in the first median position.	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/defining factors (similar to a "cued" reflection).
Balasanjeevi and Surapaneni	India	2024	Quantitative/ Qualitative	Solving 30 multiple-choice questions extracted from a textbook in Respiratory Medicine	GPT-3.5, GPT-4	Correctness: GPT-3.5: 21/30, GPT-4: 24/30.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Bonetti *et al.*	Italy	2023	Quantitative and qualitative	Solving multiple-choice questions for the Italian Residency Admission National Exam and explaining the answer	GPT-3	GPT 3 answered 87% (122/140) of the questions correctly. Two incorrect answers were due to a logical error. Explanations of the correct answers were all evaluated as appropriate.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Cai *et al.*	USA	2023	Quantitative and qualitative	Solvig exam questions used by medical residents to prepare for certification in Ophthalmology	GPT-3.5, GPT-4, Bing	GPT-4's performance in the test was similar to that of humans. However, it performed less well in questions requiring image interpretation and multiple-step diagnosis. The prevalence of hallucinations for GPT-4 was 18%.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student. However, providing inaccurate explanations may limit use.

Authors	Country	Year published	Research method	Outcomes/ Questions asked	Chatbot	Main results	Relevance for clinical reasoning
				(Basic and Clinical Sciences)			
D'Souza *et al.*	India	2023	Qualitative	Solving a hundred psychiatry clinical vignettes from a textbook	GPT-3.5	Test Performance and Preparation (grade A: 61, B: 31, C:8) The authors refined results by categories, which included clinical reasoning and ethical reasoning.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Fonseca *et al.*	Portugal	2024	Quantitative	Solving 188 questions from the American Academy of Neurology's Question of the Day app.	GPT-3.5	Score: 85% "AI chatbot provided an adequate explanation for its correct answers in 123 out of 128 cases (96.1%)."	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Hirosawa *et al.*	Japan	2023	Quantitative	Solving clinical vignettes in Internal Medicine designed for medical students and junior residents	GPT-3.5	The correct diagnosis was present in 28 out of 30 "top 10 diagnosis" lists provided by GPT and was the top diagnosis in 16. The consistency of the generation of differential diagnosis was 70.5%.	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/ defining factors (similar to a "cued" reflection).
Hudon *et al.*	Canada	2024	Quantitative/ Qualitative	Crafting Scripts Generation Tests (SCT)	GPT-3.5	"Participants could not identify which SCT was created by ChatGPT from those created by experts in the field."	Reduction of teacher workload when preparing clinical reasoning assessments. Interaction for active learning can aid students in refining their learning. Self-learning tool.
Kanjee, Crowe and Rodman	USA	2023	Quantitative	Solving cases from NEJM Clinicopathologic conferences	GPT-4	Diagnostic accuracy was 39%. The mean length of the differential diagnosis list was 9. The correct diagnosis was present in 64% of	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/

Authors	Country	Year published	Research method	Outcomes/ Questions asked	Chatbot	Main results	Relevance for clinical reasoning
						the generated lists (median position 2.5); the mean quality of the differential diagnosis (accuracy/utility) was 4.2 (in a scale of 0 to 5 proposed by the authors).	defining factors (similar to a "cued" reflection).
Kiyak and Emekli	Turkey	2024	Qualitative	Crafting Scripts Generation Tests (SCT) aimed at undergraduate medical students, with a focus on providing a diagnosis	GPT-4, GPT-4o, Claude 3, Llama 3	"Generated SCT items appear promising, but they have some limitations, such as the absence of a detailed Likert scale description."	Reduction of teacher workload when preparing clinical reasoning assessments.
Klang *et al.*	Israel	2023	Quantitative and qualitative	Writing a multiple-choice question medical examination	GPT-4	Exam quality, reviewed by specialists, showed that adjustments were needed for 15% of the questions.	Reduction of teacher workload when preparing clinical reasoning assessments. Interaction for active learning can aid students in refining their learning.
Koga, Martin and Dickson	USA	2023	Quantitative	Solving cases from Clinicopathologic conferences of neurodegenerative diseases	GPT-3.5, GPT-4, Google Bard	Diagnostic accuracy was 52% for GPT-4, 40% for Bard, and 32% for GPT3.5.	Differential diagnoses and explanations can be helpful in the development of illness scripts.
Kung *et al.*	USA	2023	Quantitative and qualitative	Solving questions from USMLE (excluded questions	GPT-3	With indeterminate responses censored/included, ChatGPT accuracy for USMLE Step 1 was 75.0%/45.4%; for Step 2CK, it was	"Partial ability to teach medicine by surfacing novel and nonobvious concepts that may not be in learners' sphere of awareness."

Authors	Country	Year published	Research method	Outcomes/ Questions asked	Chatbot	Main results	Relevance for clinical reasoning
				with images or other visual data)		61.5%/ 54.1%; and for Step 3, it was 68.8%/61.5%. At least one significant insight (novelty, nonobviousness) was present in approximately 90% of outputs.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Madrid-García *et al.*	Spain	2023	Quantitative/ Qualitative	Solving 143 rheumatology questions from the examination required for entry into specialty medical training in Spain and justifying the answer	ChatGPT, GPT-4	Score: GPT-4 93,71%, chatGPT 66,43%. "The potential usefulness of this tool, particularly in creating educational content, albeit under expert supervision." "Extreme caution should be exercised when using these models as teaching aids."	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Scherr *et al.*	USA	2023	Qualitative	Refining prompts to the simulation (related to advanced cardiac life support and intensive care), comprising a stepwise approach, user interaction responsiveness, and feedback.	GPT-3.5	Prompt one produced two desirable simulations, and 4 failed simulations that either gave incorrect feedback or did not delay feedback, while prompt two produced one desirable simulation and one failed simulation.	Reduction of teacher workload when preparing clinical reasoning assessments. Interaction for active learning can aid students in refining their learning. Self-learning tool.
Shea *et al.*	Hong Kong	2023	Quantitative	Solving real clinical cases admitted to a Geriatric ward	GPT-4	The diagnostic accuracy of GPT-4 was 67% (4/6).	It can increase confidence in diagnosis and alert "missing diagnosis". A generated list of differential diagnoses

Authors	Country	Year published	Research method	Outcomes/ Questions asked	Chatbot	Main results	Relevance for clinical reasoning
						The correct diagnosis was present in five out of six lists of Differential Diagnosis generated by GPT-4.	can help guide the student about diseases on which they should focus their study of discriminating/ defining factors (similar to a "cued" reflection).
Shieh *et al.*	USA	2024	Quantitative	Solving 109 Step 2 (Clinical Knowledge) from USMLE	GPT -3.5, GPT-4.0	Score: GPT-3.5 47,7%, GPT-4.0 87,2%. ChatGPT 4.0 accurately created a shortlist of differential diagnoses in 74.6% of the 63 case reports.	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/ defining factors (similar to a "cued" reflection)
Strong *et al.*	USA	2023	Quantitative and qualitative	Solving clinical case questions used for clinical reasoning assessment of first- and second-year medical students	GPT-3.5, GPT-4	GPT4 scored higher than students in clinical reasoning skills analysis that involved providing a case summary, a problem list, a diagnostic schema, a differential diagnosis list, and an illness script.	The summaries provided can be used as a cued reflection. AI's capacity to describe its reasoning can help students refine illness scripts and diagnostic schemas.
Wong *et al.*	USA	2024	Qualitative	Crafting unique clinical cases to be used with first-year graduate medical students	GPT-3.0	"Faculty felt that ChatGPT provided fairly accurate medical statements but could not "clinically reason" or build complexity."	Reduction of teacher workload when preparing clinical reasoning assessments.
Xie *et al.*	Australia	2023	Qualitative	Solving simulated clinical cases created by the authors, presented in steps with prompts containing clinical information	GPT-4, Bard, BingAI	Responses were graded for readability using "a combination of the Flesch Reading Ease Score, the Flesch-Kincaid Grade Level, and the Coleman-Liau Index," and also suitability using the DISCERN	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.

Authors	Country	Year published	Research method	Outcomes/ Questions asked	Chatbot	Main results	Relevance for clinical reasoning
				gathered through the viewpoint of a junior doctor.		score (for assessing the responses' quality, relevance, and equitable distribution of information).	
Yiu and Lam	UK	2023	Quantitative/ Qualitative	Solving questions from "a mock paper consisting of 300 questions deemed representative of the examination was taken from a popular question bank used widely when preparing for the Royal College of Surgeons examination."	GPT-4, Bard	The test performance of GPT was 85,7% without justification and 84,3% with forced justification. A qualitative analysis showed that LLMs might not distinguish between current and outdated guidance in their training datasets, and there were instances where responses displayed clinically inaccurate justifications.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.

REFERENCES

ARKSEY, Hilary; O'MALLEY, Lisa. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, Milton Park, v. 8, n. 1, p. 19-32, 23 Feb. 2007. DOI: https://doi.org/10.1080/1364557032000119616. Available at: https://www.tandfonline.com/doi/abs/10.1080/1364557032000119616. Accessed on: 10 Dec. 2025.

BALAS, Michael; ING, Edsel B. Conversational AI models for ophthalmic diagnosis: comparison of chatGPT and the isabel pro differential diagnosis generator. *JFO Open Ophthalmology*, Issy-les-Moulineaux, v. 1, p. 100005, 4 Mar. 2023. DOI: https://doi.org/10.1016/j.jfop.2023.100005. Available at: https://www.sciencedirect.com/science/article/pii/S2949889923000053?via%3Dihub. Accessed on: 10 Dec. 2025.

BALASANJEEVI, Gayathri; SURAPANENI, Krishna Mohan. Comparison of chatGPT version 3.5 & 4 for utility in respiratory medicine education using clinical case scenarios. *Respiratory Medicine and Research*, Issy-les-Moulineaux, v. 85, p. 101091, Jun. 2024. DOI: https://doi.org/10.1016/j.resmer.2024.101091. Available at: https://www.sciencedirect.com/science/article/pii/S2590041224000084?via%3Dihub. Accessed on: 10 Dec. 2025.

BONETTI, Mario Alessandri; GIORGINO, Riccardo; AFFLITTO, Gabriele Gallo; LORENZI, Francesca De; EGRO, Francesco M. How does chatGPT perform on the Italian residency admission national exam compared to 15,869 medical graduates? *Annals of Biomedical Engineering*, Berlin, v. 52, n. 4, p. 745-749, Apr. 2024. Available at: https://pubmed.ncbi.nlm.nih.gov/37490183/. Accessed on: 10 Dec. 2025.

BOWEN, Judith L. Educational strategies to promote clinical diagnostic reasoning. *New England Journal of Medicine*, Walthan, v. 355, n. 21, p. 2217-2225, 23 Nov. 2006. DOI: https://doi.org/10.1056/NEJMra054782. Available at: https://www.nejm.org/doi/10.1056/NEJMra054782. Accessed on: 10 Dec. 2025.

CAI, Louis Z.; SHAHEEN, Abdulla; JIN, Andrew; FUKUI, Riya; YI, Jonathan S.; YANNUZZI, Nicolas; ALABIAD, Chrisfouad. Performance of generative large language models on ophthalmology board-style questions. *American Journal of Ophthalmology*, *[S.l.]*, v. 254, p. 141-149, Oct. 2023. Doi: https://doi.org/10.1016/j.ajo.2023.05.024. Available at: https://www.sciencedirect.com/science/article/pii/S0002939423002301?via%3Dihub. Accessed on: 10 Dec. 2025.

CIVANER, M. Murat; UNCU, Yeşim; BULUT, Filiz; CHALIL, Esra Giounous; TATLI, Abdülhamit. Artificial intelligence in medical education: a cross-sectional needs assessment. *BMC medical education*, *[S.l.]*, v. 22, n. 1, p. 1-9, 9 Nov. 2022. DOI: https://doi.org/10.1186/s12909-022-03852-3. Available at: https://link.springer.com/article/10.1186/s12909-022-03852-3. Accessed on: 10 Dec. 2025.

COOPER, Nicola; BARTLETT, Maggie; GAY, Simon; HAMMOND, Anna; LILLICRAP, Mark; MATTHAN, Joanna; SINGH, Mini. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. *Medical Teacher*, *[S.l.]*, v. 43, n. 2, p. 152-159, 1 Feb. 2021. DOI: https://doi.org/10.1080/0142159X.2020.1842343. Available at: https://www.tandfonline.com/doi/full/10.1080/0142159X.2020.1842343. Accessed on: 10 Dec. 2025.

CUTRER, William B.; SULLIVAN, William M.; FLEMING, Amy E. Educational strategies for improving clinical reasoning. *Current Problems in Pediatric and Adolescent Health Care, [S.l.]*, v. 43, n. 9, p. 248-257, Oct. 2013. DOI: https://doi.org/10.1016/j.cppeds.2013.07.005. Available at: https://www.sciencedirect.com/science/article/pii/S1538544213000941?via%3Dihub. Accessed on: 10 Dec. 2025.

D'SOUZA, Russel Franco; AMANULLAH, Shabbir; MATHEW, Mary; SURAPANENI, Krishna Mohan. Appraising the performance of chatGPT in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry*, *[S.l.]*, v. 89, p. 103770, Nov. 2023. DOI: https://doi.org/10.1016/j.ajp.2023.103770. Available at: https://www.sciencedirect.com/science/article/pii/S187620182300326X?via%3Dihub. Accessed on: 10 Dec. 2025.

ERICSSON, Karl Anders. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, *[S.l.]*, v. 79, n. 10, p. S70-S81, Oct. 2004. Available at: https://pubmed.ncbi.nlm.nih.gov/15383395/#full-view-affiliation-1. Accessed on: 10 Dec. 2025.

EVA, Kevin W. What every teacher needs to know about clinical reasoning. *Medical Education*, *[S.l.]*, v. 39, n. 1, p. 98-106, Jan. 2005. DOI: https://doi.org/10.1111/j.1365-2929.2004.01972.x. Available at: https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/j.1365-2929.2004.01972.x. Accessed on: 10 Dec. 2025.

FONSECA, Ângelo; FERREIRA, Axel; RIBEIRO, Luís; MOREIRA, Sandra; DUQUE, Cristina. Embracing the future — is artificial intelligence already better? A comparative study of artificial intelligence performance in diagnostic accuracy and decision-making. *European Journal of Neurology*, *[S.l.]*, v. 31, n. 4, p. e16195, Apr. 2024. DOI: https://doi.org/10.1111/ene.16195. Available at: https://onlinelibrary.wiley.com/doi/10.1111/ene.16195. Accessed on: 10 Dec. 2025.

GORDON, Morris; DANIEL, Michelle; AJIBOYE, Aderonke; URAIBY, Hussein; XU, Nicole Y.; BARTLETT, Rangana; HANSON, Janice; HAAS, Mary; SPADAFORE, Maxwell; GRAFTON-CLARKE, Ciaran; GASIEA, Rayhan Yousef; MICHIE, Colin; CORRAL, Janet; KWAN, Brian; DOLMANS, Diana; THAMMASITBOON, Satid. A scoping review of artificial intelligence in medical education. *Medical Teacher*, *[S.l.]*, n. 84, p. 1-25, 29 Feb. 2024. DOI: https://doi.org/10.1080/0142159X.2024.2314198. Available at:

https://www.tandfonline.com/doi/full/10.1080/0142159X.2024.2314198. Accessed on: 10 Dec. 2025.

GRUNHUT, Joel; WYATT, Adam T. M.; MARQUES, Oge. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. *Journal of Medical Education and Curricular Development, [S.l.]*, v. 8, 6 Sep. 2021. DOI: https://doi.org/10.1177/23821205211036836. Available at: https://journals.sagepub.com/doi/10.1177/23821205211036836. Accessed on: 10 Dec. 2025.

HIROSAWA, Takanobu; HARADA, Yukinori; YOKOSE, Masashi; SAKAMOTO, Tetsu; KAWAMURA, Ren; SHIMIZU, Taro. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *International Journal of Environmental Research and Public Health*, *[S.l.]*, v. 20, n. 4, p. 3378, 15 Feb. 2023. DOI: https://doi.org/10.3390/ijerph20043378. Available at: https://www.mdpi.com/1660-4601/20/4/3378. Accessed on: 10 Dec. 2025.

HUDON, Alexandre; KIEPURA, Barnabé; PELLETIER, Myriam; PHAN, Véronique. Using chatGPT in psychiatry to design script concordance tests in undergraduate medical education: mixed methods study. *JMIR Medical Education*, *[S.l.]*, v. 10, p. e54067-e54067, 4 Apr. 2024. Available at: https://www.mdpi.com/1660-4601/20/4/3378. Accessed on: 10 Dec. 2025.

IBIAPINA, Cassio; MAMEDE, Sílvia; MOURA, Alexandre; ELÓI-SANTOS, Silvana; GOG, Tamara van. Effects of free, cued and modelled reflection on medical students' diagnostic competence. *Medical Education*, *[S.l.]*, v. 48, n. 8, p. 796-805, Aug. 2014. DOI: https://doi.org/10.1111/medu.12435. Available at: https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/medu.12435. Accessed on: 10 Dec. 2025.

KANJEE, Zahir; CROWE, Byron; RODMAN, Adam. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, *[S.l.]*, v. 330, n. 1, p. 78, 3 Jul. 2023. Available at: https://jamanetwork.com/journals/jama/fullarticle/2806457. Accessed on: 10 Dec. 2025.

KIYAK, Yavuz Selim; EMEKLI, Emre. A prompt for generating script concordance test using chatGPT, claude, and llama large language model chatbots. *Revista Española de Educación Médica*, Murcia, v. 5, n. 3, 15 May 2024. DOI: https://doi.org/10.6018/edumed.612381. Available at: https://revistas.um.es/edumed/article/view/612381. Accessed on: 10 Dec. 2025.

KLANG, Eyal; PORTUGEZ, Shir; GROSS, Raz; KASSIF LERNER, Reut; BRENNER, Alina; GILBOA, Mayan; ORTAL, Tal; RON, Sophi; ROBINZON, Vered; MEIRI, Hila; SEGAL, Gad. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. *BMC Medical Education*, *[S.l.]*, v. 23, n. 1, p. 772, 17 Oct.

2023. Available at: https://link.springer.com/article/10.1186/s12909-023-04752-w. Accessed on: 10 Dec. 2025.

KOGA, Shunsuke; MARTIN, Nicholas B.; DICKSON, Dennis W. Evaluating the performance of large language models: chatGPT and google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathology*, *[S.l.]*, p. e13207, 8 Aug. 2023. DOI: https://doi.org/10.1111/bpa.13207. Available at: https://onlinelibrary.wiley.com/doi/10.1111/bpa.13207. Accessed on: 10 Dec. 2025.

KUNG, Tiffany H.; CHEATHAM, Morgan; MEDENILLA, Arielle; SILLOS, Czarina; LEON, Lorie De; ELEPAÑO, Camille; MADRIAGA, Maria; AGGABAO, Rimel; DIAZ-CANDIDO, Giezel; MANINGO, James; TSENG, Victor. Performance of chatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *[S.l.]*, v. 2, n. 2, p. e0000198, 9 Feb. 2023. DOI: https://doi.org/10.1371/journal.pdig.0000198. Available at: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000198. Accessed on: 10 Dec. 2025.

LEE, Hyunsu. The rise of chatGPT: exploring its potential in medical education. *Anatomical Sciences Education*, *[S.l.]*, v. 17, n. 5, 14 Mar. 2023. DOI: https://doi.org/10.1002/ase.2270. Available at: https://anatomypubs.onlinelibrary.wiley.com/doi/10.1002/ase.2270. Accessed on: 10 Dec. 2025.

MADRID-GARCÍA, Alfredo; ROSALES-ROSADO, Zulema; FREITES-NUÑEZ, Dalifer; PÉREZ-SANCRISTÓBAL, Inés; PATO-COUR, Esperanza; PLASENCIA-RODRÍGUEZ, Chamaida; CABEZA-OSORIO, Luis; ABASOLO-ALCÁZAR, Lydia; LÉON-MATEOS, Leticia; FERNÁNDEZ-GUTIÉRREZ, Benjamín; RODRÍGUEZ-RODRÍGUEZ, Luis. Harnessing chatGPT and GPT-4 for evaluating the rheumatology questions of the spanish access exam to specialized medical training. *Scientific Reports*, *[S.l.]*, v. 13, n. 1, p. 22129, 13 Dec. 2023. DOI: https://doi.org/10.1038/s41598-023-49483-6. Available at: https://www.nature.com/articles/s41598-023-49483-6. Accessed on: 10 Dec. 2025.

MAMEDE, Sílvia. What does research on clinical reasoning have to say to clinical teachers? *Scientia Medica,* Porto Alegre, v. 30, p. 1-8, 15 Jul. 2020. DOI: https://doi.org/10.15448/1980-6108.2020.1.37350. Available at: https://revistaseletronicas.pucrs.br/scientiamedica/article/view/37350. Accessed on: 10 Dec. 2025.

MAMEDE, Sílvia; SCHMIDT, Henk G. Deliberate reflection and clinical reasoning: founding ideas and empirical findings. *Medical Education*, *[S.l.]*, v. 57, n. 1, p. 76-85, Jan. 2023. DOI: https://doi.org/10.1111/medu.14863. Available at: https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/medu.14863. Accessed on: 10 Dec. 2025.

PRAKASH, Shivesh; SLADEK, Ruth M.; SCHUWIRTH, Lambert. Interventions to improve diagnostic decision making: a systematic review and meta-analysis on reflective strategies. *Medical Teacher, [S.l.]*, v. 41, n. 5, p. 517-524, 4 May 2019. DOI:

https://doi.org/10.1080/0142159X.2018.1497786. Available at: https://www.tandfonline.com/doi/full/10.1080/0142159X.2018.1497786. Accessed on: 10 Dec. 2025.

PREIKSAITIS, Carl; ROSE, Christian. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Medical Education*, *[S.l.]*, v. 9, n. 1, p. e48785, 2023. DOI: https://doi.org/10.2196/48785. Available at: https://mededu.jmir.org/2023/1/e48785. Accessed on: 10 Dec. 2025.

SCHERR, Riley; HALASEH, Faris F.; SPINA, Aidin; ANDALIB, Saman; RIVERA, Ronald. ChatGPT interactive medical simulations for early clinical education: case study. *JMIR Medical Education*, *[S.l.]*, v. 9, p. e49877, 10 Nov. 2023. DOI: https://doi.org/10.2196/49877. Available at: https://mededu.jmir.org/2023/1/e49877. Accessed on: 10 Dec. 2025.

SCHMIDT, Hank G.; RIKERS, Remy. M. J. P. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education, [S.l.]*, v. 41, n. 12, p. 1133-1139, 14 Nov. 2007. DOI: https://doi.org/10.1111/j.1365-2923.2007.02915.x. Available at: https://asmepublications.onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2007.02915.x. Accessed on: 10 Dec. 2025.

SHEA, Yat-Fung; LEE, Cynthia Min Yao; IP, Whitney Chin Tung; LUK, Dik Wai Anderson; WONG, Stephanie Sze Wing. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. *JAMA Network Open, [S.l.]*, v. 6, n. 8, p. e2325000, 14 Aug. 2023. DOI: https://doi.org/10.1001/jamanetworkopen.2023.25000. Available at: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2808251. Accessed on: 10 Dec. 2025.

SHIEH, Allen; TRAN, Brandon; HE, Gene; KUMAR, Mudit; FREED, Jason A.; MAJETY, Priyanka. Assessing chatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Scientific Reports, [S.l.]*, v. 14, n. 1, p. 9330, 23 Apr. 2024. DOI: https://doi.org/10.1038/s41598-024-58760-x. Available at: https://pubmed.ncbi.nlm.nih.gov/38654011/. Accessed on: 10 Dec. 2025.

STRONG, Eric; DIGIAMMARINO, Alicia; WENG, Yingjie; KUMAR, Andre; HOSOMANI, Poonam; HOM, Jason; CHEN, Jonathan H. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Internal Medicine, [S.l.]*, v. 183, n. 9, p. 1028-1030, 1 Sep. 2023. DOI: https://doi.org/10.1001/jamainternmed.2023.2909. Available at: https://pubmed.ncbi.nlm.nih.gov/37459090/. Accessed on: 10 Dec. 2025.

WONG, Kristin; FAYNGERSH, Alla; TRABA, Christin; CENNIMO, David; KOTHARI, Neil; CHEN, Sophia. Using chatGPT in the development of clinical reasoning cases: a qualitative study. *Cureus*, *[S.l.]*, v. 16, n. 5, 31 May 2024. DOI: https://doi.org/10.7759/cureus.61438. Available at: https://www.cureus.com/articles/253053-using-chatgpt-in-the-development-of-clinical-reasoning-cases-a-qualitative-study#!/. Accessed on: 10 Dec. 2025.

WOODS, Nicole N.; NEVILLE, Alan J.; LEVINSON, Anthony J.; HOWEY, Elizabeth H. A.; OCZKOWSKI, Wieslaw J.; NORMAN, Geoffrey R. The value of basic science in clinical diagnosis. *Academic Medicine*, *[S.l.]*, v. 81, n. 10, p. S124-S127, Oct. 2006. DOI: https://doi.org/10.1097/00001888-200610001-00031. Available at: https://pubmed.ncbi.nlm.nih.gov/17001122/. Accessed on: 10 Dec. 2025.

XIE, Yi; SETH, Ishith; HUNTER-SMITH, David J.; ROZEN, Warren M.; SEIFMAN, Marc A. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. *ANZ Journal of Surgery*, *[S.l.]*, v. 94, n. 1-2, p. 68-77, Aug. 2023. DOI: https://doi.org/10.1111/ans.18666. Available at: https://onlinelibrary.wiley.com/doi/10.1111/ans.18666. Accessed on: 10 Dec. 2025.

YIU, Allen; LAM, Kyle. Performance of large language models at the MRCS part A: a tool for medical education? *The Annals of the Royal College of Surgeons of England*, *[S.l.]*, v. 107, n. 6, p. 434-440, Dec. 2023. DOI: https://doi.org/10.1308/rcsann.2023.0085. Available at: https://publishing.rcseng.ac.uk/doi/10.1308/rcsann.2023.0085. Accessed on: 10 Dec. 2025.

APPENDIX 1

Search strategies were adapted to the indexing structure of each database. This approach was designed to maximize sensitivity and retrieve gray literature relevant to the research question in each database.

WoS: (((((((((TI=(chatGPT)) OR TI=(chatbot)) OR TI= (large language model) OR TI= (artificial intelligence)) AND TI= (medical education)) OR TI= (medical school)) OR TI= (medical student)) OR TI= (clinical education)) AND ALL= (clinical reasoning)) OR ALL= (clinical judgment)) AND ALL= (clinical decision-making);

Google Scholar: (chatGPT, OR chatbot, OR "artificial intelligence") AND ("medical education OR "medical student") AND ("clinical reasoning" OR "clinical decision-making");

PubMed: ((((chatgpt) OR (chatbot)) OR (artificial intelligence)) AND (medical education)) AND (clinical reasoning).

Guilherme Freitas bernardo Ferreira

Guilherme Freitas Bernardo Ferreira is a neurologist with a master's degree in Health Education/Clinical Reasoning Education. He served as a professor at Unifenas-BH from 2021 to 2025.

gfreitasbernardo@gmail.com

Alexandre Sampaio Moura

Alexandre Sampaio Moura is an infectious diseases physician working as a full professor at the Graduate Program in Medicine and Biomedicine at Faculdade Santa Casa, Belo Horizonte, Brazil. Prof. Moura conducts research in medical education, with a particular interest in clinical reasoning and competence-based assessment.

alexandremoura@faculdadesantacasabh.edu.br

Lígia Maria Cayres Ribeiro

Ligia Cayres Ribeiro is an internal medicine physician with a PhD in Clinical Reasoning. She is a researcher at the University Medical Center Groningen, where she investigates how technology can enhance evidence-informed educational practices.

l.m.cayres.ribeiro@umcg.nl

Maria Aparecida Turci

Maria Aparecida Turci is a public health professional working as a full professor at the Graduate Program in Medicine at Professor Edson Antônio Velano University, Belo Horizonte, Brazil. Prof. Turci conducts research in public health and health professions education.

mariaturci@gmail.com

Sílvia Mamede

Sílvia Mamede is a guest professor at the Wenckebach Institute (WIOO), Lifelong Learning, Education and Assessment Research Network (LEARN), University Medical Center Groningen, Netherlands. She conducts research on clinical reasoning and diagnostic error in medicine; educational strategies for the teaching of clinical reasoning; reflection and experiential learning in medical education and clinical practice.

s.mamede@erasmusmc.nl

Como citar este documento – ABNT

FERREIRA, Guilherme Freitas Bernardo; MOURA, Alexandre Sampaio; RIBEIRO, Lígia Maria Cayres; TURCI, Maria Aparecida; MAMEDE, Sílvia. Como a inteligência artificial generativa pode facilitar o ensino do raciocínio clínico: uma revisão de escopo. *Revista Docência do Ensino Superior*, Belo Horizonte, v. 15, e058339, p. 1-27, 2025. DOI: https://doi.org/10.35699/2237-5864.2025.58339.