# COMO OS GRANDES MODELOS DE LINGUAGEM (LLMS) ESTÃO IMPACTANDO A EDUCAÇÃO MÉDICA? UMA REVISÃO SISTEMÁTICA

Mariana Serra Duarte (IC) e Carlos Augusto Gottsfritz Fernandes (Orientador)

**Apoio: PIVIC Mackenzie**

## RESUMO

Objetivo: mapear as aplicações, os desfechos educacionais e as barreiras associadas ao uso de Large Language Models (LLMs) no ensino médico na literatura de 2022 a 2025. Método: revisão sistemática registrada no PROSPERO (CRD420261445760) e conduzida conforme o PRISMA 2020. Foram realizadas buscas nas bases PubMed/MEDLINE e Embase, atualizadas em 25 de junho de 2026. Incluíram-se estudos empíricos primários envolvendo estudantes de medicina, residentes, médicos em formação continuada ou docentes que utilizaram LLMs em atividades educacionais. A triagem e a extração foram conduzidas por dois revisores, com resolução das divergências por consenso. Como os estudos eram muito diferentes entre si, levando em conta delineamentos, populações e instrumentos, realizou-se síntese narrativa, sem metanálise. Resultados: foram incluídos 30 estudos, conduzidos em 15 países, com maior concentração na China, Reino Unido, Alemanha e Estados Unidos. Predominaram estudos experimentais, quase-experimentais e transversais, além de métodos mistos e estudos prospectivos. Os modelos da família GPT foram os mais utilizados, e pacientes virtuais, simulação clínica, treinamento de anamnese e raciocínio clínico constituíram as aplicações mais frequentes. Os achados apontaram benefícios potenciais em desempenho, eficiência, satisfação e apoio ao raciocínio, porém, os resultados variaram bastante. Entre as principais barreiras destacaram-se limitações metodológicas, risco de informações incorretas, limitações técnicas e necessidade de supervisão humana. Não foi realizada avaliação formal do risco de viés dos estudos incluídos. Conclusão: os LLMs apresentam potencial para a educação médica, mas sua adoção requer avaliação crítica, validação pedagógica e investigação de efeitos de longo prazo.

**Palavras-chave:** Inteligência Artificial. Educação Médica. Revisão Sistemática.

## ABSTRACT

Objective: to map the applications, educational outcomes, and barriers associated with the use of Large Language Models (LLMs) in medical education in the literature published between 2022 and 2025. Methods: this systematic review was registered in PROSPERO (CRD420261445760) and conducted in accordance with the PRISMA 2020 guidelines. Searches were performed in PubMed/MEDLINE and Embase and were last updated on June 25, 2026. Primary empirical studies involving medical students, residents, physicians in

continuing education, or faculty members who used LLMs in educational activities were included. Screening and data extraction were conducted by two reviewers, with disagreements resolved by consensus. Because the studies differed substantially in terms of design, populations, and assessment instruments, a narrative synthesis was performed without meta-analysis. Results: thirty studies conducted in 15 countries were included, with the highest concentration of studies in China, the United Kingdom, Germany, and the United States. Experimental, quasi-experimental, and cross-sectional studies predominated, alongside mixed-methods and prospective studies. GPT-family models were the most frequently used, while virtual patients, clinical simulation, history-taking training, and clinical reasoning were the most common applications. The findings suggested potential benefits in performance, efficiency, satisfaction, and reasoning support; however, the results varied considerably. The main barriers included methodological limitations, the risk of inaccurate information, technical limitations, and the need for human supervision. No formal risk-of-bias assessment of the included studies was conducted. Conclusion: LLMs show potential for medical education, but their adoption requires critical evaluation, pedagogical validation, and further investigation of long-term effects.

**Keywords:** Artificial Intelligence. Medical Education. Systematic Review

## INTRODUÇÃO

Os *Large Language Models (LLMs)* são sistemas de inteligência artificial (IA) desenvolvidos para compreender e processar a linguagem humana. Esses modelos são treinados com grandes volumes de dados textuais, permitindo diversas tarefas, como geração de resumos e textos, tradução e análise de informações (Ribeiro; Tsunoda, 2025). Na área do ensino médico, o uso de LLMs constitui uma questão relevante devido às suas barreiras, desafios e limitações na formação integral dos futuros profissionais da saúde. Hirani *et al.* (2024) destacam que não houve uma data ou ano específico em que a IA foi integrada na saúde, mas sim uma evolução gradual desde os anos 70, com o primeiro consultor médico artificial, o *INTERNIST-1*, criado em 1971. Sua criação representou uma grande mudança no uso dessa tecnologia em pesquisas clínicas, pois apresentava um potencial significativo de aliviar parte do diagnóstico clínico que recaía sobre os profissionais de saúde e oferecia aos médicos um mecanismo para conferir seus diagnósticos diferenciais. Dessa forma, com o passar do tempo o uso de inteligências artificiais foi cada vez mais se disseminando por diversas áreas, entrando também no ensino médico, principalmente nos últimos 3 anos com o uso de LLMs generativas.

A IA abrange uma gama de tecnologias que conseguem executar funções ligadas a habilidades humanas, como aprendizado, identificação de padrões e decisões. Dentro desse universo, a IA generativa é o termo usado para designar sistemas que geram novos conteúdos, enquanto os LLMs são modelos dedicados ao processamento e à geração da linguagem natural, sendo o ChatGPT um dos exemplos mais conhecidos dessa tecnologia (Boscardin *et al.*, 2024). A popularização dessas ferramentas foi significativa após a liberação do ChatGPT para o público em novembro de 2022, o que ampliou seu acesso e fomentou discussões sobre suas potencialidades, limitações e aplicações na educação médica.

Segundo Lindbäck *et al.* (2025), deve haver o questionamento do uso indiscriminado da IA, visto que acelerar a transformação de fatos em conhecimento adquirido, pode comprometer a aprendizagem de conhecimentos ditos essenciais, na medida em que os estudantes recorrem a “atalhos” no ensino. Com isso, faz-se necessária uma análise sistemática de benefícios e limitações do uso da IA, principalmente de LLMs, no ensino médico, uma vez que esse ensino compreende a formação de graduação, pós-graduação e educação continuada de médicos, capacitando-os com conhecimentos essenciais, habilidades clínicas e competências éticas.

Diante disso, a importância dessa revisão sistemática é reforçada pela grande demanda observada na rotina e nas salas de aula, incluindo o uso de plataformas de LLMs por estudantes para sintetização, resumos de artigos, personalização dos estudos e por oferecerem ao ser humano um suporte tecnológico para aumentar e melhorar a produtividade.

A partir disso, o estudo poderá contribuir para consolidar os dados disponíveis e fornecer fundamentos para futuras pesquisas e para o aprimoramento do ensino médico baseado em IA.

Dessa forma, com a rápida expansão e a variedade de aplicações de LLMs neste contexto, torna-se essencial organizar e apresentar o conhecimento científico disponível. A pergunta de pesquisa pode ser definida como: Quais são as aplicações, os desfechos educacionais e as barreiras associadas ao uso de LLMs no ensino médico conforme reportado na literatura científica recente? O objetivo deste trabalho foi identificar e categorizar as aplicações de LLMs no ensino médico reportadas na literatura; sintetizar os desfechos educacionais associados ao uso de LLMs, incluindo desempenho acadêmico, percepções de estudantes e impactos no processo de aprendizagem; e listar as barreiras, desafios e limitações relacionadas à implementação de LLMs no ensino médico (Quadro 1). A revisão seguiu uma metodologia sistemática de busca e seleção de artigos científicos recentes, utilizando o PubMed e Embase, repertórios online internacionais e mundialmente reconhecidos como base de dados para buscas, de forma a mapear o cenário atual de utilização de LLMs no ensino médico nos últimos 3 anos (2022-2025).

Quadro 1: Estratégia PICO

| **Elemento** | **Definição desta revisão** |
|---|---|
| **P** (População) | Estudantes de medicina (graduação), residentes, médicos em educação médica continuada e docentes |
| **I** (Intervenção/Exposição) | *Large Language Models (LLMs)* |
| **C** (Comparação) | Não aplicável, pois esta será uma revisão descritiva / exploratória. Este trabalho realizará um mapeamento abrangente do uso de LLMs na educação médica |
| **O** (Desfechos) | Desfechos educacionais (desempenho, retenção, satisfação) e barreiras de implementação |

Fonte: Autoria própria (2026)

As revisões existentes restringem-se ao ChatGPT, deixando de fora modelos como DeepSeek e Kimi Chat, que esta revisão mostra serem centrais na produção chinesa; outras produções são também predominantemente narrativas, sem protocolo registrado nem *checklist* PRISMA; incluem estudos de desempenho isolado do modelo em exames; e são limitadas a produções predominantemente da Ásia, Europa e América do Norte, assim, não incluindo diferenças culturais e regionais do uso dos LLMs no ensino médico.

## DESENVOLVIMENTO DO ARGUMENTO

A revisão seguiu as etapas definidas para revisões sistemáticas e foi registrada na base PROSPERO sob o número CRD420261445760. Também foi guiada pelas recomendações do *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*

(PRISMA) (Page *et al.*, 2021). O objetivo foi garantir transparência e reprodutibilidade no processo de busca, seleção e análise dos estudos. Os autores declaram não haver conflitos de interesse.

Foram consideradas publicações disponíveis cuja primeira publicação editorial online tenha ocorrido entre novembro de 2022 e 31 de dezembro de 2025[1]. A busca foi feita nas bases de dados PubMed/MEDLINE e Embase. Para encontrar os estudos, foram utilizados descritores e termos livres ligados aos conceitos de inteligência artificial, *Large Language Models* e educação médica, assim, combinando esses termos usando os operadores booleanos AND e OR. A estratégia de busca foi ajustada para cada base de dados. Por fim, identificamos estudos que poderiam ser relevantes, 556 registros (PubMed/MEDLINE, n=204; Embase, n=352).

Ademais, com o uso do software Rayyan, foi feita a remoção de duplicatas (8) e a triagem dos estudos (548) conforme os critérios definidos. Foram incluídos estudos que tivessem como população estudantes de medicina em nível de graduação, residentes, médicos em formação continuada ou professores de medicina. Os estudos deveriam avaliar como intervenção ou exposição o uso de *Large Language Models*, incluindo ferramentas como ChatGPT, Gemini, Claude, NotebookLM, LLaMA e Med-PaLM, entre outros modelos desse tipo. Além disso, os trabalhos deveriam estar em contexto educacional, envolvendo ensino, aprendizagem ou treinamento, sem que o uso da tecnologia ficasse restrito à prática clínica assistencial.

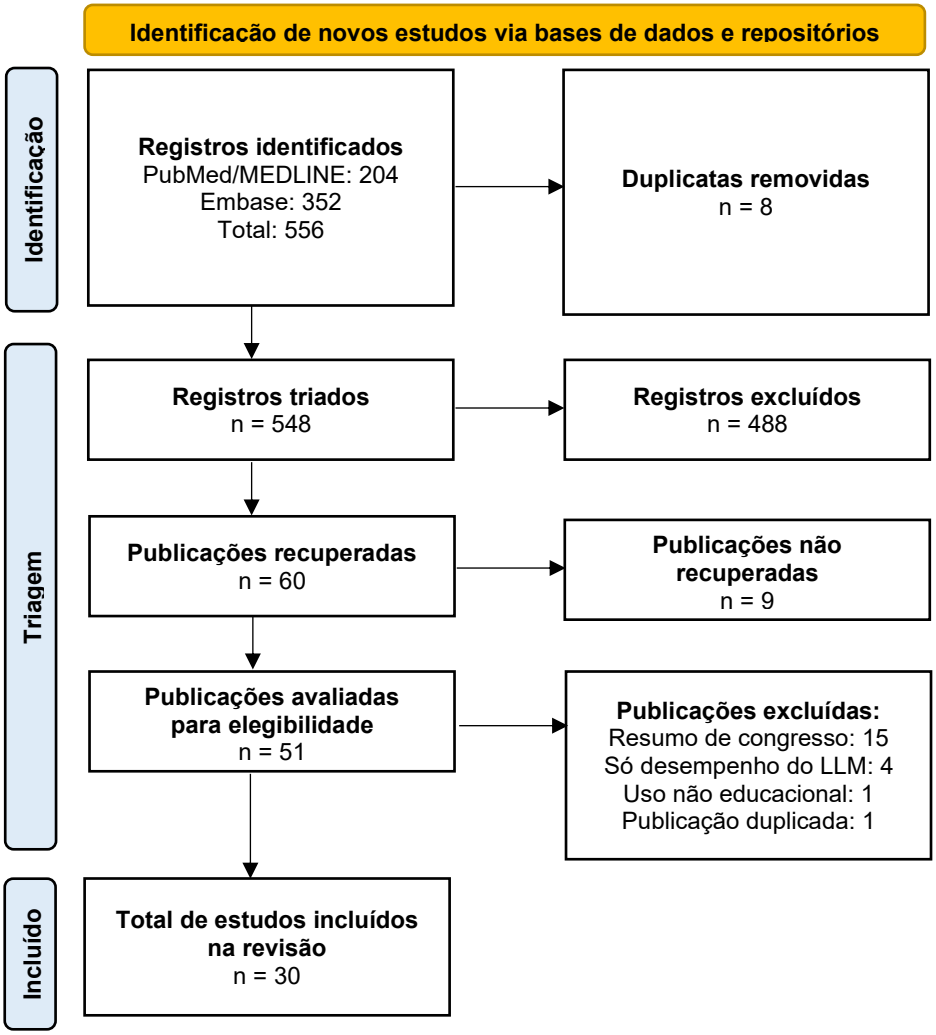
No entanto, foram excluídas revisões de literatura, incluindo revisões narrativas, sistemáticas e de escopo. Também ficaram de fora editoriais, cartas ao editor e comentários. Estudos que abordavam formas de inteligência artificial que não fossem *Large Language Models*, como métodos tradicionais de *machine learning*, *deep learning* aplicado à análise de imagens ou redes neurais convolucionais, também foram excluídos. Trabalhos que avaliavam só o uso clínico assistencial de LLMs, sem parte educacional, não foram considerados. Da mesma forma, estudos que analisavam só o desempenho do próprio LLM em exames médicos, sem participação de estudantes ou profissionais em formação usando a ferramenta no processo educacional, foram excluídos. Estudos feitos só em outras áreas da saúde, como enfermagem ou fisioterapia, também foram excluídos quando não incluíam estudantes de medicina. Não foram considerados resumos de conferências sem apresentação de dados completos, protocolos de pesquisa sem resultados, capítulos de livros e *pré-prints*. Por fim, foram excluídas publicações listadas como retratadas nos registros bibliográficos das bases

[1] Para definir o período de publicação, foi considerada a primeira data em que o artigo ficou disponível online em sua versão editorial, mesmo que depois ele tenha sido publicado em um volume de outro ano. Assim, foram incluídos estudos disponibilizados online entre novembro de 2022 e 31 de dezembro de 2025. Esse foi o caso de Fan *et al.,* publicado online em 23 de dezembro de 2025, embora depois tenha sido incluído no volume 26 do periódico, referente a 2026.

de dados consultadas. O processo de identificação, triagem, elegibilidade e inclusão dos estudos foi apresentado em um diagrama de fluxo seguindo o guia PRISMA (Figura 1).

Dessa forma, na triagem 488 foram excluídos. Dos 60 relatos buscados para recuperação, 9 não foram obtidos apesar de busca ativa em múltiplas fontes, restando 51 avaliados quanto à elegibilidade. Destes, 21 foram excluídos: 15 por se tratar de resumos de conferência ou congresso sem dados completos, 4 por avaliarem exclusivamente o desempenho do modelo sem uso da ferramenta por aprendizes, 1 por ausência de componente educacional e 1 por constituir publicação duplicada de estudo já incluído. Trinta estudos compuseram a síntese final.

Figura 1: Diagrama de fluxo PRISMA 2020




Fonte: Autoria própria (2026)

Foram extraídos os seguintes campos: referência completa; país; desenho do estudo; população (n, nível de formação e especialidade); LLM utilizado (modelo e versão); categoria de aplicação; descrição da intervenção; comparador; desfechos educacionais avaliados e respectivos instrumentos; principais resultados quantitativos; achados sobre alucinações e acurácia; barreiras e limitações relatadas; financiamento e conflitos de interesse; e

observações do extrator. Depois, os dados foram organizados e analisados em um quadro (Quadro 2), ajudando a identificar padrões, semelhanças e diferenças entre os resultados. No entanto, a síntese se manteve narrativa, uma vez que os delineamentos, intervenções, comparação e desfechos não seguiram um padrão, assim, inviabilizando a meta-análise.

Quadro 2: Extração dos estudos incluídos na revisão (n = 30)

| Estudo | Delineamento e população | LLM e aplicação | Principais resultados | Barreiras relatadas |
|---|---|---|---|---|
| Paciente virtual simulado e treino de anamnese, comunicação e raciocínio clínico (n = 11) | | | | |
| **1. Holderried *et al.* (2024)** Alemanha | Prospectivo não controlado 106 estudantes de medicina (93% no 3º ano) | GPT-4 (API) Paciente virtual + feedback automatizado | Completude da anamnese e concordância com avaliador humano. Kappa geral de 0,832; respostas plausíveis em mais de 99% dos casos; 8 das 45 categorias de feedback com kappa inferior a 0,6. | Um único modelo, um único prompt e um único caso clínico; critério binário; sem medida de habilidade real. |
| **2. Brügge *et al.* (2024)** Alemanha | ECR duplo-cego, 2 braços 21 estudantes de medicina | ChatGPT 3.5 Paciente virtual + feedback estruturado | Escore CRI-HTI. Grupo com feedback superou o controle após 4 sessões (3,60 vs. 3,02; $p = 0,049$; $\eta^2p = 0,198$); sem ganho no domínio "focar perguntas" ($p = 0,265$). | Amostra pequena; conduzido em alemão; estudantes adaptaram a fala para contornar a latência do modelo. |
| **3. Hirosawa *et al.* (2025)** Japão | ECR cruzado, piloto 20 médicos residentes (PGY1–2) | GPTs (ChatGPT customizado) Paciente virtual para treino de entrevista médica | Escala Likert de 6 pontos em 6 domínios. Estação de IA inferior à tradicional (4,89 vs. 5,47; $p < 0,001$); sem diferença significativa em raciocínio clínico ($p = 0,10$). | Piloto sem poder estatístico; modalidade de interação distinta entre os braços (digitada vs. falada); unicêntrico. |
| **4. Guo *et al.* (2025)** Canadá | Avaliação de ferramenta com survey pós-uso 37 estudantes do 1º ano | GPT-4 + Llama-3-70B Paciente virtual interativo (voz e texto) no formato OSCE | Escala Likert de 5 pontos. Facilidade de uso 4,73; qualidade do conteúdo 4,70; feedback 4,41; avaliação favorável em relação à aula expositiva (4,14; $p < 0,0001$). | Viés de autosseleção; imagens estáticas; 8 de 37 relataram cenários pouco representativos; autores são criadores da ferramenta. |
| **5. Borg *et al.* (2025)** Suécia | Prova de conceito, métodos mistos 15 estudantes de medicina | GPT-3.5-turbo Paciente virtual em robô social (Furhat) | Questionário validado de desenho de paciente virtual. Plataforma robótica superior em autenticidade (4,5 vs. 3,9; $p = 0,04$) e em efeito de aprendizagem (4,4 vs. 4,1; $p = 0,01$). | Amostra pequena; ausência de exame físico e de exames laboratoriais; diálogo por vezes mecânico; acesso limitado ao equipamento. |

| Estudo | Delineamento e população | LLM e aplicação | Principais resultados | Barreiras relatadas |
|---|---|---|---|---|
| **6. Borg *et al.* (2024)** Suécia | Qualitativo exploratório intervencional 23 estudantes do 3º ano | GPT-3.5-turbo Paciente virtual em robô social (Furhat) | Entrevistas semiestruturadas com análise temática. Três temas identificados: autenticidade, aplicação do paciente virtual e pontos fortes e limitações; robô percebido como mais autêntico e engajador. | Ausência de exame físico; estudo não cegado; uso do inglês como segunda língua; achados não generalizáveis. |
| **7. Öncü *et al.* (2025)** Turquia | Triangulação simultânea, métodos mistos 21 médicos internos (6º ano) | ChatGPT-4o Paciente padronizado por IA para gestão de caso clínico | Autoavaliação de 35 itens comparada à observação externa. Ausência de correlação entre autoavaliação e desempenho observado, sugestiva de efeito Dunning-Kruger. | Falhas técnicas e de reconhecimento do idioma turco; pressão de tempo apontada por 10 participantes; apenas dois casos clínicos. |
| **8. Tyrrell *et al.* (2025)** Reino Unido | ECR cruzado, multicêntrico 378 estudantes do 3º ano | SimConverse (modelo não divulgado) Paciente virtual por voz para habilidades de consulta | Escala Calgary-Cambridge e custo por estudante. Ambos os braços melhoraram; IA inferior ao ator (diferença de −0,36; IC95% −0,60 a −0,13; p = 0,04); satisfação 8,09 vs. 9,21 (p < 0,001); custo de £33,48 vs. £61,75. | Desfechos autorrelatados; exposição única e de curta duração; sem avaliação somativa objetiva de desempenho. |
| **9. Jacobs *et al.* (2025)** Reino Unido | Viabilidade, métodos mistos sequencial 15 estudantes do 3º ano e 3 médicos generalistas | GPT-3.5-turbo + ElevenLabs Paciente virtual por voz para treino de comunicação | Escala ITEM abreviada. Motivação intrínseca 16,5/20 (d = 2,09); usabilidade 12,0/15 (d = 2,18); fidelidade (6,0/10) e imersão (8,5/15) apenas moderadas. | Latência de 2 a 3 segundos; ausência de pistas não verbais; linguagem técnica demais para um paciente leigo; sem grupo controle. |
| **10. Comulada *et al.* (2025)** Estados Unidos | Piloto prospectivo de braço único 10 estudantes de medicina | GPT-4 (voz) Paciente virtual e agente de feedback (protocolo SPIKES) | System Usability Scale e preparo autorrelatado. SUS de 91/100; o número de participantes que se consideravam preparados para dar más notícias passou de 2 para 10; 90% consideraram o feedback útil. | Amostra muito pequena; feedback percebido como pouco crítico; sem medida de transferência para desempenho clínico real. |
| **11. Bentegeac *et al.* (2025)** França | Piloto prospectivo multicêntrico de validação 91 estudantes de medicina (anos 4 a 6) | GPT-4o Paciente e examinador virtuais em estações de OSCE | Fidelidade do paciente virtual e concordância com padrão-ouro. Autenticidade de 98,6%; ICC de 0,94 na nota global, mas de apenas 0,44 em itens de atitude e comunicação; 91,7% avaliaram a utilidade em 7 ou mais de 10. | Baixa confiabilidade como avaliador em itens subjetivos; não simula procedimentos nem imagens; sem desfecho de aprendizagem pré e pós. |

| | | | | |
|---|---|---|---|---|
| Aprendizagem baseada em problemas, em casos e em simulação de cenário assistida por LLM (n = 3) | | | | |
| **12. Hou *et al.* (2025)** China | Quase-experimental, 2 braços 60 residentes de hematologia do 2º ano | DeepSeek V3/R1 PBL assistido por LLM | Exame padronizado e questionário de 15 itens. Grupo DeepSeek superior no escore total (86,03 vs. 81,73; p < 0,001) e em 13 dos 15 itens; sem diferença em habilidades clínicas (p = 0,078) nem em engajamento (p = 0,127). | Alocação por cronograma (viés de seleção); amostra pequena e unicêntrica; curta duração; possível efeito Hawthorne. |
| **13. Li *et al.* (2025)** China | Comparativo entre duas turmas, quase-experimental 89 estudantes de medicina clínica | Kimi Chat 2.0 CBL assistida por IA e IA como avaliadora de tarefas | Nota das tarefas, tempo e prova final. Grupo experimental mais rápido (2,6 h vs. 5,5 h) e com nota final superior (77,3 vs. 66,5; p < 0,05); aceitação da IA subiu de 5,5 para 7,6. | Redução da interação professor-aluno; homogeneização das respostas; percepção de menor estímulo ao pensamento inovador. |
| **14. Zheng *et al.* (2024)** China | Quase-experimental, métodos mistos 66 estudantes do 1º ano do ciclo clínico | ChatGPT 3.5 Simulação de cenário com roteiro e banco de questões gerados por IA | Prova teórica, Mini-CEX e escala de pensamento crítico. Grupo experimental superior na prova (82,91 vs. 72,16), no Mini-CEX (76,24 vs. 70,19) e em pensamento crítico (92,94 vs. 89,31; p < 0,001). | Amostra pequena e unicêntrica; versão do modelo já superada à época; respostas insatisfatórias a perguntas muito específicas. |
| Geração de conteúdo educacional (n = 2) | | | | |
| **15. Zhang, Q. *et al.* (2025)** China | Estudo de viabilidade com survey pré e pós 12 casos avaliados por 8 docentes; 20 estudantes de graduação | ChatGPT 4.0 Geração de casos clínicos com questões de múltipla escolha | Rubrica de qualidade de 12 itens e questionário de percepção. Escore médio de qualidade de 52,33 de 60; 95% dos estudantes concordaram que os casos enriqueceram os recursos de aprendizagem. | 16,67% dos casos apresentaram alucinações; não gera imagens clínicas reais; questões com dificuldade mal calibrada. |
| **16. Fąferek *et al.* (2025)** Consórcio iCoViP (multicêntrico europeu) | Metodológico com revisão parcialmente cegada e survey 31 estudantes de medicina; 54 sessões de paciente virtual | ChatGPT-4o Geração de pacientes virtuais para treino de raciocínio clínico | Checklist iCoViP e questionário de 15 itens. Sem diferença de qualidade, tempo ou acurácia entre pacientes virtuais humanos e gerados por IA; apenas 43,2% dos estudantes identificaram corretamente a origem. | Blueprint com padrões repetitivos e baixa diversidade sociodemográfica; escopo mais estreito; geração de imagens clínicas realistas inviável. |
| Assistente de ensino, tutor curricular e apoio à decisão clínica (n = 5) | | | | |
| **17. Fan *et al.* (2026)** China | ECR estratificado, unicêntrico 24 estudantes do 4º ano e 24 residentes | DeepSeek Assistente de ensino com base de conhecimento curada | Prova teórica e retenção em 30 dias. Ganho imediato no grupo experimental (83,21 vs. 73,08; p = 0,007), sem diferença na retenção (64,00 vs. 60,58; p = 0,277). | Amostra pequena e unicêntrica; ausência de ganho de retenção; desenho não isola qual componente gerou o efeito. |

| | | | | |
|---|---|---|---|---|
| **18. Thesen e Park (2025)** Estados Unidos | Observacional de implementação, duas coortes 92 e 98 estudantes do 2º ano | GPT-4 e GPT-4o (RAG) Tutor curricular restrito ao material do curso | Métricas de uso e utilidade percebida. 360 conversas e 2.946 mensagens em dois anos; aumento de 329,6% no volume de conversas nos 3 dias que antecedem as provas ($p < 0,001$); utilidade de 2,8 e 3,3 de 5. | Frustração com o escopo restrito (36,8% dos comentários); queda de 59,4% no uso entre as coortes; amostra autosselecionada. |
| **19. Ng *et al.* (2025)** Reino Unido | ECR cruzado, métodos mistos 20 estudantes do 1º ano | Lenny AI (GPT-4o) *Chatbot* tutor de anatomia clínica | Desempenho objetivo e 12 domínios de percepção. Sem diferença significativa de desempenho ($p = 0,065$ e $p = 0,634$); percepção significativamente favorável ao *chatbot* em 5 domínios ($p < 0,05$). | Ganhos restritos à fluência e à recuperação factual; risco de "confiança sem competência"; autores desenvolveram a ferramenta. |
| **20. Robleto *et al.* (2024)** Estados Unidos | Coorte com survey transversal 73 estudantes do 1º ano | Glass AI (GPT-4) Ferramenta de apoio diagnóstico | Questionário de 16 itens em 3 domínios. 96% relataram aumento de confiança no diagnóstico e 97% concordaram que a ferramenta conduziu ao diagnóstico correto; 43% consideraram a explicação insuficiente. | Viés de não resposta; caso clínico simples; unicêntrico; 68% manifestaram preocupação com o impacto sobre a profissão médica. |
| **21. Montagna *et al.* (2025)** Itália | Comparativo entre grupos, alocação não randomizada 16 estudantes do último ano | ChatGPT 3.5 Apoio à decisão comparado a diretrizes e repositórios | Completude e acurácia avaliadas por 4 examinadores cegos. Acurácia: diretrizes 4,0; ChatGPT 3,7; repositório online 3,3 ($p = 0,02$); pior desempenho do ChatGPT em tomada de decisão clínica (3,5 vs. 4,1). | Amostra muito pequena; ausência de master-prompting; erro crítico de interpretação laboratorial; não fornece citação de fontes. |
| Uso autodirigido e percepções sobre o uso (n = 6) | | | | |
| **22. Zeng *et al.* (2025)** China | Quase-experimental com grupo controle 142 estudantes do 3º ano | ChatGPT (versão não especificada) Revisão autodirigida de conteúdo sobre doença rara | Pré e pós-teste de 135 pontos e tempo de revisão. Sem diferença no escore total ($p = 0,28$); redução de 43% no tempo de revisão ($p < 0,0001$); desempenho superior em questões de raciocínio clínico complexo ($p < 0,05$). | Unicêntrico; sessão única, sem medida de retenção; prova centrada em recordação factual. |
| **23. Fusshöller *et al.* (2025)** Alemanha | Transversal, métodos mistos 84 estudantes de medicina | ChatGPT (versão não especificada) Uso autodirigido para estudo | Frequência de uso e percepção de impacto. 76,2% já usaram com fins educacionais; uso maior em período de provas ($p = 0,003$); 70,2% relatam economia de tempo, mas apenas 31% percebem aplicação do conhecimento e 15,5% retenção. | 73,3% preocupados com informações incorretas e falta de transparência das fontes; taxa de resposta de 26,7%. |
| **24. Alnaim *et al.* (2025)** Arábia Saudita | Transversal, três universidades públicas 386 estudantes da área da saúde | ChatGPT (versão não especificada) Uso em ambiente de aprendizagem remota | Escalas Likert em 7 domínios. 96,8% concordam que a integração da ferramenta beneficia a aprendizagem remota; maiores escores em aprendizagem personalizada (4,35) e comunicação (4,23). | Ceticismo (89,9%) e preocupações éticas (85,4%); dependência excessiva (68,9%); amostra de conveniência. |

| | | | | |
|---|---|---|---|---|
| **25. Sunmboye *et al.* (2025)** Reino Unido | Transversal, quantitativo 200 estudantes analisados, anos 1 a 5 | Não especificado Uso autodirigido para aprendizagem personalizada | AI Use Scale e percepções. Maior dependência entre estudantes do 1º ano, com queda nos anos posteriores; gênero associado a maior preocupação com desinformação (p = 0,017) e a maior adoção de estratégias de checagem (p = 0,045). | Amostragem por conveniência; dados autorrelatados; desenho transversal que impede inferência causal. |
| **26. Nwe *et al.* (2025)** Malásia | Transversal, descritivo 184 estudantes pré-clínicos de MBBS | ChatGPT (56,4%), Gemini (27,0%), Copilot (12,5%) Uso autodirigido para estudo | Familiaridade e efetividade percebida em 4 domínios. 100% já usaram alguma ferramenta e 72,8% usam diariamente; 87,5% consideram a IA efetiva para gerar novas ideias e 66,8% para pensamento crítico. | 83,7% preocupados com dependência excessiva e 78,8% com perda de pensamento crítico; unicêntrico; dados autorrelatados. |
| **27. Abu Hammour *et al.* (2024)** Jordânia | Transversal, amostragem por conveniência 244 estudantes do 4º ano ou superior | ChatGPT (versão não especificada) Uso autodirigido no estudo e no treino clínico | Percepções, práticas e preocupações. 88,5% conheciam a ferramenta e 86,9% concordaram que estudantes podem se beneficiar; apenas 32,7% concordaram que auxilia decisões de tratamento. | 78,3% apontam risco de erros e imprecisões; preocupações com privacidade do paciente e com acesso desigual à ferramenta. |
| Outras aplicações (n = 3) | | | | |
| **28. Kumar *et al.* (2025)** Índia | Quase-experimental intra-sujeito 36 residentes de ortopedia | ChatGPT (versão não especificada) Auxílio à escrita científica | Rubrica de 6 competências avaliada por painel de 3 pesquisadores. Ganho em técnicas de escrita (2,7 para 4,4) e em uso da linguagem (2,6 para 4,6; p < 0,001); sem ganho em identificação de problemas de pesquisa (p = 0,210) ou em compreensão de métodos (p = 0,080). | Conteúdo bem estruturado, porém genérico e sem profundidade crítica; risco de fabricação e plágio; unicêntrico e sem grupo controle. |
| **29. Zhang, L. *et al.* (2025)** China | ECR com randomização simples 186 médicos de hospitais terciários e não terciários | ChatGPT 4.0 Treinamento breve para manejo clínico em formação continuada | Teste de competência em asma pediátrica (0 a 100). Ganho de cerca de 20 pontos no grupo intervenção; uso regular da ferramenta subiu de 6,3% para 62% em duas semanas. | Seguimento curto; incentivos financeiros aos participantes; instrumento de teste sem validação psicométrica formal. |
| **30. Naamati-Schneider e Alt (2025)** Israel | Quase-experimental intra-sujeito 80 graduandos em gestão de sistemas de saúde | ChatGPT e plataforma Poe (GPT-3.5-Turbo) Alfabetização em IA e desenvolvimento de *chatbot* | Escala de prompt-crafting e questionário de motivação, com modelagem por equações estruturais. Criar um *chatbot* associou-se a maior motivação intrínseca que usar o ChatGPT (β = 0,75; p < 0,001); sem diferença nas habilidades de prompting entre as fases. | População fora da formação médica; medidas inteiramente autorrelatadas; ordem fixa das atividades; ausência de dados longitudinais. |

Fonte: Autoria própria (2026)

Com essa análise, é perceptível que os 30 estudos incluídos foram conduzidos predominantemente na China (n=7), Reino Unido (n=4), Alemanha (n=3) e Estados Unidos (n=3), com estudos adicionais no Japão, Canadá, Suécia, Turquia, Itália, Índia, França, Israel,

Jordânia, Malásia, Arábia Saudita e um consórcio multicêntrico europeu (iCoViP). Essa distribuição geográfica é um achado relevante, visto que há uma concentração expressiva de produção chinesa, frequentemente usando alternativas ao ChatGPT (DeepSeek, Kimi Chat) por restrições de acesso locais.

Quanto ao desenho dos estudos, a amostra é heterogênea: 5 ensaios clínicos randomizados (ECR), 7 estudos quase-experimentais com grupo comparador não randomizado, 3 estudos prospectivos de braço único sem controle, 6 estudos transversais tipo survey de percepção/uso, e um conjunto de estudos de métodos mistos ou qualitativos com forte componente exploratório. Essa mistura de desenhos é a principal razão pela qual uma metanálise quantitativa direta dificilmente seria viável; a síntese abaixo é, portanto, narrativa.

Já o modelo de linguagem predominante é da família GPT (GPT-3.5/ChatGPT 3.5, GPT-4 e GPT-4o), presente em cerca de 20 dos 30 estudos, seguido por plataformas proprietárias construídas sobre esses modelos (Glass AI, OSCEai, SimConverse, Lenny AI, ECOSBot, NeuroBot/getcody.ai) e por alternativas chinesas (DeepSeek; Kimi Chat).

Com isso, foi possível identificar também as aplicações dos LLMs no ensino médico, sendo elas cinco grandes categorias funcionais: paciente virtual simulado e treino de anamnese, comunicação e raciocínio clínico, a categoria mais numerosa, reunindo praticamente metade dos estudos, incluindo desde *chatbots* baseados em texto até simulações por voz e robôs sociais *embodied*; aprendizagem baseada em problemas/casos (PBL) assistida por LLM, em que o modelo atua como facilitador ou fonte de apoio durante discussões em grupo; geração de conteúdo educacional, casos clínicos com questões de múltipla escolha, design de pacientes virtuais, roteiros de vídeo e bancos de questões; assistente de ensino e tutor curricular, incluindo plataformas completas com interface e ferramentas de avaliação para conectar modelos de linguagem a uma bases de dados restritas ao material do curso e ferramentas de apoio à decisão clínica comparadas a diretrizes e bases de referência tradicionais; uso autodirigido e não controlado pelo próprio estudante, captado por *surveys* de frequência, finalidade e percepção de uso, sem intervenção desenhada pelos pesquisadores. Houve também um caso adicional que trata do uso de LLM para apoiar escrita científica de residentes, e outro, do breve treinamento de médicos (não estudantes) para uso clínico do ChatGPT.

Igualmente, uma grande parte dos estudos quase-experimentais, juntamente com ensaios clínicos randomizados demonstraram efeitos favoráveis ao uso de LLMs sobre desfechos relacionados a melhora do conhecimento e do desempenho acadêmico. No caso do ensino baseado em problemas, Hou *et al.* (2025) viram que alunos que usaram o DeepSeek tiveram um desempenho melhor que o grupo tradicional, com notas superiores em quase todos os itens avaliados. Porém, essa melhora não apareceu nas habilidades práticas ou no engajamento do grupo. O mesmo aconteceu com Zheng *et al.* (2024) em simulações

de cardiologia, na qual o grupo com ChatGPT foi melhor na teoria e no pensamento crítico. Li *et al.* (2025) também notaram que o uso do Kimi Chat ajudou os alunos a terem notas melhores e a terminarem as tarefas mais rápido. Já no treinamento de anestesiologia, Fan *et al.* (2026) perceberam um aumento no conhecimento logo após o uso do DeepSeek, mas esse efeito sumiu após 30 dias no novo teste aplicado, o que mostra que o aprendizado imediato nem sempre vira memória de longo prazo. Por fim, Zhang *et al.* (2025) mostraram que o ChatGPT ajudou médicos a ganharem competência sobre asma pediátrica, e que o uso da ferramenta teve um aumento expressivo ao longo de duas semanas, de 6,3% para 62%.

Além da influência sobre o desempenho, os estudos sugerem que os benefícios dos LLMs podem ser relevantes para tarefas que necessitam de maior complexidade cognitiva. No estudo de Zeng *et al.* (2025), embora não tenha sido observada diferença estatisticamente significativa no escore total entre os grupos com e sem ChatGPT, os participantes que utilizaram a ferramenta apresentaram redução de 43% no tempo de revisão e desempenho superior especificamente em questões que demandavam raciocínio clínico complexo, integração diagnóstica e síntese multidisciplinar. Esse resultado converge com os achados de Hou *et al.* (2025), nos quais a superioridade do grupo assistido por DeepSeek também se concentrou em determinados domínios da avaliação, não sendo observada de maneira uniforme em todos os componentes analisados. Em conjunto, esses achados indicam que o impacto educacional dos LLMs pode estar menos relacionado à simples recuperação de informações factuais e mais associado ao suporte oferecido em tarefas que exigem grande integração de conhecimentos, raciocínio clínico complexo e resolução de problemas.

Em conjunto, os estudos também identificaram situações em que os LLMs apresentaram desempenho inferior ou semelhante aos métodos tradicionais. Na avaliação de entrevistas médicas, Hirosawa *et al.* (2025) observaram menor desempenho da estação de IA em relação à realizada com ator humano, especialmente em comunicação, anamnese, exame físico e manejo, sem diferença apenas no raciocínio clínico. De forma semelhante, Tyrrell *et al.* (2025) verificaram menor ganho de desempenho e satisfação com a simulação por IA em comparação à simulação com atores, embora com custo significativamente menor. Já Montagna *et al.* (2025) identificaram menor acurácia do ChatGPT em comparação às diretrizes clínicas, incluindo um erro crítico na interpretação de exames laboratoriais.

Por outro lado, alguns estudos demonstraram equivalência entre LLMs e métodos tradicionais. Fąferek *et al.* (2025) não encontraram diferenças na qualidade, tempo ou acurácia diagnóstica entre pacientes virtuais gerados por IA e por clínicos. Da mesma forma, Ng *et al.* (2025) não observaram diferenças no desempenho objetivo com o Lenny AI, embora os participantes tenham apresentado maior satisfação, percepção de qualidade, facilidade de uso e engajamento com o *chatbot*. Assim, foi percebido que o uso dessa ferramenta pode não trazer o melhor desempenho, mas melhora a percepção subjetiva dos participantes.

Os resultados dessa revisão confirmam estudos anteriores que mostram o potencial dos LLMs como ferramentas para ajudar na educação médica, principalmente em atividades como aprendizagem personalizada, resolução de casos complexos e desenvolvimento do raciocínio clínico. No entanto, esses resultados trazem um ponto importante ao mostrar que os benefícios são mais claros em tarefas que precisam de integração de conhecimentos, raciocínio clínico e uso eficiente do tempo de estudo. Eles aparecem muito menos em atividades que só pedem recuperação de fatos ou que dependem de habilidades práticas e comunicação, como em simulações. Esse padrão também mostra que um desempenho melhor em provas teóricas não garante mais competência clínica. Assim, como outros trabalhos já apontaram, os LLMs parecem ser mais úteis como ferramentas que apoiam o aprendizado. Eles não devem substituir o ensino tradicional, o professor ou a prática.

## CONSIDERAÇÕES FINAIS

Em relação ao primeiro objetivo dessa revisão, identificar e categorizar as aplicações dos LLMs no ensino médico, a literatura analisada demonstrou uma utilização muito diversa dessas ferramentas. Os LLMs foram utilizados como apoio para aprendizagem, resolução e discussão de casos clínicos, simulações, preparação para avaliações, tutoria personalizada, desenvolvimento do raciocínio clínico complexo e fornecimento de feedback. Esses achados evidenciam que sua aplicação não se restringe à obtenção de informações, podendo atuar em diferentes etapas da aprendizagem.

Além disso, quanto ao segundo objetivo, sintetizar os desfechos educacionais associados ao uso de LLMs, os resultados indicaram benefícios principalmente relacionados ao desempenho acadêmico, ao conhecimento, ao raciocínio clínico e à eficiência do processo de aprendizagem. Entretanto, esses efeitos não foram uniformes; alguns estudos demonstraram superioridade dos grupos assistidos pela IA, enquanto outros encontraram desempenho equivalente ou inferior com métodos tradicionais. Dessa forma, a principal conclusão da revisão é que os LLMs apresentam potencial como ferramentas de apoio à aprendizagem médica, principalmente em tarefas que exigem uma integração de conhecimentos, raciocínio clínico e resolução de problemas, mas não devem substituir os métodos tradicionais de ensino ou da interação com professores e pacientes na graduação e no ensino continuado.

Já em relação à busca de barreiras, desafios e limitações relacionadas à implementação dos LLMs, foram identificadas preocupações quanto à precisão e a acurácia das informações, possibilidade de respostas incorretas, dependência excessiva da tecnologia, questões éticas, privacidade de dados e limitações para o desenvolvimento e avaliação de habilidades práticas e comunicativas. A diversidade dos métodos utilizados nos estudos, dos modelos avaliados e dos desfechos analisados dificulta a comparação direta dos resultados.

Essas limitações reforçam os achados de revisões anteriores, que apontam a necessidade de maior rigor metodológico e de estudos que avaliem efetivamente os impactos educacionais dos LLMs, e não apenas seu desempenho e/ou potencial teórico.

Para pesquisas futuras, recomenda-se a realização de estudos multicêntricos, com amostras maiores, grupos de comparação adequados e acompanhamento longitudinal e indicadores padronizados, permitindo avaliar não apenas o desempenho imediatamente após a intervenção, mas também a retenção do conhecimento e a aplicação da aprendizagem para situações clínicas. Também são necessários estudos que avaliem diferentes modelos de LLMs e diferentes aplicações pedagógicas, além de investigações sobre aspectos éticos, segurança, privacidade e formação dos docentes e estudantes com o uso crítico dessas ferramentas. Destaca-se ainda a escassez de evidências da América Latina. A literatura disponível permanece fortemente concentrada em países de alta renda e regiões como América do Norte, Ásia e Europa, o que limita a compreensão de como fatores culturais, educacionais locais, linguísticos e de infraestrutura podem influenciar a implementação dos LLMs em diferentes contextos. Assim, a produção de estudos latino-americanos representa uma importante lacuna e uma oportunidade para desenvolver evidências mais contextualizadas às necessidades e características dos sistemas de educação médica da região. Houve uma limitação pela ausência de uma avaliação formal da qualidade metodológica dos estudos incluídos, o que dificultou determinar o grau de confiança nos resultados. Além disso, a diversidade de métodos e instrumentos utilizados nos estudos reforça a importância dessa avaliação.

**Declaração de uso de inteligência artificial generativa:** foram utilizados os modelos ChatGPT (OpenAI) e Claude (Anthropic) como ferramentas auxiliares de formatação e extração preliminar de informações. Não foram utilizadas ferramentas de inteligência artificial para realizar a triagem ou decidir sobre a inclusão e exclusão dos estudos. Todo o conteúdo produzido ou extraído com auxílio dessas ferramentas foi revisado e validado pelos autores com base nas fontes originais, que assumem integral responsabilidade pelo conteúdo apresentado.

## REFERÊNCIAS

ABU HAMMOUR, A.; ABU HAMMOUR, K.; ALHAMAD, H.; NASSAR, R.; EL-DAHIYAT, F.; SAWAQED, M.; ALLAN, A.; MANASEER, Q.; ABU HAMMOUR, M.; HALBOUP, A.; ABU FARHA, R. Exploring Jordanian medical students' perceptions and concerns about ChatGPT in medical education: a cross-sectional study. **Journal of Pharmaceutical Policy and Practice**, [London], v. 17, n. 1, 2429000, 25 nov. 2024. DOI: https://doi.org/10.1080/20523211.2024.2429000. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39600801/. Acesso em: 25 jul. 2026.

ALNAIM, N.; ALSANAD, D. S.; ALBELALI, S.; ALMULHEM, M.; ALMUHANNA, A. F.; ATTAR, R. W.; ALSAHLI, M.; ALBAGMI, S.; BAKHSHWAIN, A. M.; ALMAZROU, S.; ALMUTAIRI, S. A.; ABOALSAMH, H. M.; ARIF, W. M.; ALSADHAN, A. A.; ALSEDRAH, I. T.; ALANEZI, F.; ALIBRAHIM, D.; ALQAHTANI, N. S. Effectiveness of ChatGPT in remote learning environments: An empirical study with medical students in Saudi Arabia. **Nutrition and Health**, [London], v. 31, n. 3, p. 1035-1046, set. 2025. DOI: https://doi.org/10.1177/02601060241273596. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39150345/. Acesso em: 25 jul. 2026.

BENTEGEAC, R.; FLORENS, N.; MAANAOUI, M.; MAISONS, V.; LANOT, A.; BOBOT, M.; BRILLAND, B.; GLOWACKI, F.; GÉRARD, E.; HAZZAN, M.; AMOUYEL, P.; LE GUELLEC, B.; HAMROUN, A. ECOSBot: a multicenter validation pilot study of a generative AI tool for OSCE-based nephrology training. **Clinical Kidney Journal**, [Oxford], v. 18, n. 10, sfaf308, 6 out. 2025. DOI: https://doi.org/10.1093/ckj/sfaf308. Disponível em: https://academic.oup.com/ckj/article/18/10/sfaf308/8275750. Acesso em: 25 jul. 2026.

BORG, A.; GEORG, C.; JOBS, B.; HUSS, V.; WALDENLIND, K.; RUIZ, M.; EDELBRING, S.; SKANTZE, G.; PARODIS, I. Virtual Patient Simulations Using Social Robotics Combined With Large Language Models for Clinical Reasoning Training in Medical Education: Mixed Methods Study. **Journal of Medical Internet Research**, [Toronto], v. 27, e63312, 3 mar. 2025. DOI: https://doi.org/10.2196/63312. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40053778/. Acesso em: 25 jul. 2026.

BORG, A.; JOBS, B.; HUSS, V.; GENTLINE, C.; ESPINOSA, F.; RUIZ, M.; EDELBRING, S.; GEORG, C.; SKANTZE, G.; PARODIS, I. Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. **Rheumatology International**, [Berlin], v. 44, n. 12, p. 3041-3051, dez. 2024. DOI: https://doi.org/10.1007/s00296-024-05731-0. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39412574/. Acesso em: 25 jul. 2026.

BOSCARDIN, C. K.; GIN, B.; GOLDE, P. B.; HAUER, K. E. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. **Academic Medicine**, [S. l.], v. 99, n. 1, p. 22-27, 2024. DOI: https://doi.org/10.1097/ACM.0000000000005439. Disponível em: https://academic.oup.com/academicmedicine/article/99/1/22/8343954. Acesso em: 14 maio 2026.

BRÜGGE, E.; RICCHIZZI, S.; ARENBECK, M.; KELLER, M. N.; SCHUR, L.; STUMMER, W.; HOLLING, M.; LU, M. H.; DARICI, D. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. **BMC Medical Education**, [London], v. 24, n. 1, 1391, 28 nov. 2024. DOI: https://doi.org/10.1186/s12909-024-06399-7. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39609823/. Acesso em: 25 jul. 2026.

COMULADA, W. S.; GANZ, P. A.; HUANG, Y. M.; WEISMAN, D.; SUGARMAN, A.; LEE, D.; SHAH, S.; GELBERG, L. A Pilot Test of an AI Voice-Driven Simulation With Feedback for Medical Students to Practice Discussing Diagnostic Mammogram Results With Patients. **Cureus**, [Palo Alto, CA], v. 17, n. 10, e95606, out. 2025. DOI: https://doi.org/10.7759/cureus.95606. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41322766/. Acesso em: 25 jul. 2026.

FĄFEREK, J.; KONONOWICZ, A. A.; BOGUTSKA, N.; DA SILVA DOMINGUES, V.; DAVYDOVA, N.; FRANKOWSKA, A.; IGUACEL, I.; MAYER, A.; MORIN, L.; PAVLYUKOVICH, N.; POPOVA, I.; SHCHUDROVA, T.; SUDACKA, M.; SZYDLAK, R.; HEGE, I. Applying ChatGPT to plan and create a realistic collection of virtual patients for clinical reasoning training. **BMC Medical Education**, [London], v. 25, n. 1, 1277, 2 out. 2025.

DOI: https://doi.org/10.1186/s12909-025-08006-9. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41039567/. Acesso em: 25 jul. 2026.

FAN, H.; DU, W.; YAN, J.; JIANG, H. Application of DeepSeek-based AI teaching assistant in teaching anesthesiology theories. **BMC Medical Education**, [London], v. 26, n. 1, 138, 2026. DOI: https://doi.org/10.1186/s12909-025-08494-9. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41437042/. Acesso em: 25 jul. 2026.

FUSSHÖLLER, A.; LECHNER, F.; SCHLICKER, N.; MUEHLENSIEPEN, F.; MAYR, A.; KUHN, S.; HIRSCH, M. C.; KNITZA, J. Perceptions, Usage, and Educational Impact of ChatGPT Among Medical Students in Germany: Cross-Sectional Mixed Methods Survey. **JMIR Formative Research**, [Toronto], v. 9, e81484, 11 nov. 2025. DOI: https://doi.org/10.2196/81484. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41218187/. Acesso em: 25 jul. 2026.

GUO, E.; RAMCHANDANI, R.; PARK, Y. J.; GUPTA, M. OSCEai: personalized interactive learning for undergraduate medical education. **Canadian Medical Education Journal**, [Saskatoon, Saskatchewan], v. 16, n. 6, p. 7-14, dez. 2025. DOI: https://doi.org/10.36834/cmej.79220. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41584941/. Acesso em: 25 jul. 2026.

HIRANI, R.; NORUZI, K.; KHURAM, H.; HUSSAINI, A. S.; AIFUWA, E. I.; ELY, K. E.; LEWIS, J. M.; GABR, A. E.; SMILEY, A.; TIWARI, R. K.; ETIENNE, M. Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. **Life**, [S. l.], v. 14, n. 5, 557, 2024. DOI: https://doi.org/10.3390/life14050557. Disponível em: https://www.mdpi.com/2075-1729/14/5/557. Acesso em: 08 maio 2026.

HIROSAWA, T.; YOKOSE, M.; SAKAMOTO, T.; HARADA, Y.; TOKUMASU, K.; MIZUTA, K.; SHIMIZU, T. Utility of Generative Artificial Intelligence for Japanese Medical Interview Training: Randomized Crossover Pilot Study. **JMIR Medical Education**, [Toronto], v. 11, e77332, 1 ago. 2025. DOI: https://doi.org/10.2196/77332. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40749190/. Acesso em: 25 jul. 2026.

HOLDERRIED, F.; STEGEMANN-PHILIPPS, C.; HERRMANN-WERNER, A.; FESTL-WIETEK, T.; HOLDERRIED, M.; EICKHOFF, C.; MAHLING, M. A Language Model-Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. **JMIR Medical Education**, [Toronto], v. 10, e59213, 16 ago. 2024. DOI: https://doi.org/10.2196/59213. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39150749/. Acesso em: 25 jul. 2026.

HOU, J.; AN, F.; QIN, H.; ZHANG, L.; WANG, J.; ZHANG, C.; FAN, D. Application of DeepSeek-assisted problem-based learning in hematology residency training. **BMC Medical Education**, [London], v. 25, n. 1, 1291, 2 out. 2025. DOI: https://doi.org/10.1186/s12909-025-07852-x. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41039507/. Acesso em: 25 jul. 2026.

JACOBS, C.; JOHNSON, H.; TAN, N.; BROWNLIE, K.; JOINER, R.; THOMPSON, T. Application of AI Communication Training Tools in Medical Undergraduate Education: Mixed Methods Feasibility Study Within a Primary Care Context. **JMIR Medical Education**, [Toronto], v. 11, e70766, 24 out. 2025. DOI: https://doi.org/10.2196/70766. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41135055/. Acesso em: 25 jul. 2026.

KUMAR, R.; P, A. K.; SELVAM, A. A.; DHAMU, I.; BALASUBRAMANIAN, N. The impact of ChatGPT on orthopedic residents' scientific writing: A quasi-experimental study. **Journal of Clinical Orthopaedics and Trauma**, [Delhi], v. 69, 103116, out. 2025. DOI:

https://doi.org/10.1016/j.jcot.2025.103116. Disponível em: https://www.journal-cot.com/article/S0976-5662(25)00214-0/fulltext. Acesso em: 25 jul. 2026.

LI, L.; ZHANG, W.; ZHANG, K.; YANG, Y.; WANG, L.; ZUO, L.; SUN, Y.; PENG, Q. The role of generative AI tools in case-based learning and teaching evaluation of medical biochemistry. **BMC Medical Education**, [London], v. 25, n. 1, 1185, 22 ago. 2025. DOI: https://doi.org/10.1186/s12909-025-07567-z. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40847414/. Acesso em: 25 jul. 2026.

LINDBÄCK, Y. et al. Generative artificial intelligence in physiotherapy education: great potential amidst challenges – a qualitative interview study. **BMC Medical Education**, [S. l.], v. 25, 603, 2025. DOI: https://doi.org/10.1186/s12909-025-07106-w. Disponível em: https://doi.org/10.1186/s12909-025-07106-w. Acesso em: 25 jul. 2026.

MONTAGNA, M.; CHIABRANDO, F.; DE LORENZO, R.; ROVERE QUERINI, P. Impact of Clinical Decision Support Systems on Medical Students' Case-Solving Performance: Comparison Study with a Focus Group. **JMIR Medical Education**, [Toronto], v. 11, e55709, 18 mar. 2025. DOI: https://doi.org/10.2196/55709. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40101183/. Acesso em: 25 jul. 2026.

NAAMATI-SCHNEIDER, L.; ALT, D. Developing AI literacy in medical and health professions education: Exploring ChatGPT, *Chatbot*s development, and simulation-based learning. **Nurse Education in Practice**, [S. l.], v. 88, 104572, out. 2025. DOI: https://doi.org/10.1016/j.nepr.2025.104572. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41027189/. Acesso em: 25 jul. 2026.

NG, I. S. H.; SIU, A.; HAN, C. S. J.; HO, O. S. H.; SUN, J.; MARKIV, A.; KNIGHT, S.; SAGOO, M. G. Evaluating a Custom *Chatbot* in Undergraduate Medical Education: Randomised Crossover Mixed-Methods Evaluation of Performance, Utility, and Perceptions. **Behavioral Sciences**, [Basel], v. 15, n. 9, 1284, 19 set. 2025. DOI: https://doi.org/10.3390/bs15091284. Disponível em: https://www.mdpi.com/2076-328X/15/9/1284. Acesso em: 25 jul. 2026.

NWE, T. M.; SHAMSHOL, N. A. H.; JAAFAR, N. N.; SYAHRUDDIN, S. H.; YUSAINI, T. A.; SAHA'ARIF, T. F. N. T.; LWIN, S.; YEE, K. T.; YI, M. S.; LATT, S. S. The Role of Artificial intelligence in Medical training, with its Applicability, Efficiency, Potential, and Challenges among Preclinical Students. **Research Journal of Pharmacy and Technology**, [Raipur], v. 18, n. 6, p. 2508-2516, 2025. DOI: https://doi.org/10.52711/0974-360X.2025.00358. Disponível em: https://www.rjptonline.org/AbstractView.aspx?PID=2025-18-6-11. Acesso em: 25 jul. 2026.

ÖNCÜ, S.; TORUN, F.; ÜLKÜ, H. H. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. **BMC Medical Education**, [London], v. 25, n. 1, 278, 20 fev. 2025. DOI: https://doi.org/10.1186/s12909-025-06877-6. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39979969/. Acesso em: 25 jul. 2026.

PAGE, M. J. et al. A declaração PRISMA 2020: diretriz atualizada para relatar revisões sistemáticas. **Epidemiologia e Serviços de Saúde**, Brasília, DF, v. 31, n. 2, 2022. DOI: https://doi.org/10.1590/S1679-49742022000200033. Disponível em: http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742022000201700&lng=pt&nrm=iso. Acesso em: 21 abr. 2026.

RIBEIRO, P. F. R.; TSUNODA, D. F. Geração de textos dinâmicos e criativos com Grandes Modelos de Linguagem (LLMs): uma revisão sistemática. **Texto Livre**, Belo Horizonte, v. 18, e60103, 2025. DOI: https://doi.org/10.1590/1983-3652.2025.60103. Disponível em:

https://www.scielo.br/j/tl/a/VyBtxKGrV686cYT7GsQ9kQJ/?lang=pt. Acesso em: 08 maio 2026.

ROBLETO, E.; HABASHI, A.; KAPLAN, M.-A. B.; RILEY, R. L.; ZHANG, C.; BIANCHI, L.; SHEHADEH, L. A. Medical students' perceptions of an artificial intelligence (AI) assisted diagnosing program. **Medical Teacher**, [London], v. 46, n. 9, p. 1180-1186, set. 2024. DOI: https://doi.org/10.1080/0142159X.2024.2305369. Disponível em: https://pubmed.ncbi.nlm.nih.gov/38306667/. Acesso em: 25 jul. 2026.

SUNMBOYE, K.; STRAFFORD, H.; NOORESTANI, S.; WILISON-PIRIE, M. Exploring the influence of artificial intelligence integration on personalized learning: a cross-sectional study of undergraduate medical students in the United Kingdom. **BMC Medical Education**, [London], v. 25, n. 1, 570, 17 abr. 2025. DOI: https://doi.org/10.1186/s12909-025-07084-z. Disponível em: https://pubmed.ncbi.nlm.nih.gov/40247306/. Acesso em: 25 jul. 2026.

THESEN, T.; PARK, S. H. A generative AI teaching assistant for personalized learning in medical education. **npj Digital Medicine**, [London], v. 8, n. 1, 627, 4 nov. 2025. DOI: https://doi.org/10.1038/s41746-025-02022-1. Disponível em: https://www.nature.com/articles/s41746-025-02022-1. Acesso em: 25 jul. 2026.

TYRRELL, E. G.; SANDHU, S. K.; BERRY, K.; GHANNAM, S. F.; LEWIS, S. A.; CROWFOOT, D.; SAHOTA, G. S.; CARSON, J.; WILSON, E. E.; TAGGAR, J. Web-Based AI-Driven Virtual Patient Simulator Versus Actor-Based Simulation for Teaching Consultation Skills: Multicenter Randomized Crossover Study. **JMIR Formative Research**, [Toronto], v. 9, e71667, 20 nov. 2025. DOI: https://doi.org/10.2196/71667. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41264856/. Acesso em: 25 jul. 2026.

XU, T.; WENG, H.; LIU, F.; YANG, L.; LUO, Y.; DING, Z.; WANG, Q. Current Status of ChatGPT Use in Medical Education: Potentials, Challenges, and Strategies. **Journal of Medical Internet Research**, [S. l.], v. 26, e57896, 2024. DOI: https://doi.org/10.2196/57896. Disponível em: https://doi.org/10.2196/57896. Acesso em: 03 jul. 2026.

ZENG, J.; SUN, K.; QIN, P.; LIU, S. Enhancing ophthalmology students' awareness of retinitis pigmentosa: assessing the efficacy of ChatGPT in AI-assisted teaching of rare diseases—a quasi-experimental study. **Frontiers in Medicine**, [Lausanne], v. 12, 1534294, 18 mar. 2025. DOI: https://doi.org/10.3389/fmed.2025.1534294. Disponível em: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1534294/full. Acesso em: 25 jul. 2026.

ZHANG, L.; YANG, G.; YUAN, J.; YUAN, S.; ZHANG, J.; CHEN, J.; TANG, M.; ZHANG, Y.; LIN, J.; ZHAO, L.; YIN, Y. Enhancing pediatric asthma management in underdeveloped regions through ChatGPT training for doctors: a randomized controlled trial. **Frontiers in Pediatrics**, [Lausanne], v. 13, 1519751, 3 jul. 2025. DOI: https://doi.org/10.3389/fped.2025.1519751. Disponível em: https://www.frontiersin.org/journals/pediatrics/articles/10.3389/fped.2025.1519751/full. Acesso em: 25 jul. 2026.

ZHANG, Q.; HUANG, Z.; HUANG, Y.; WANG, G.; ZHANG, R.; YANG, J.; CHENG, Y.; CHEN, B.; WANG, H.; QIU, K.; CHEN, H. Generative AI in medical education: feasibility and educational value of LLM-generated clinical cases with MCQs. **BMC Medical Education**, [London], v. 25, n. 1, 1502, 27 out. 2025. DOI: https://doi.org/10.1186/s12909-025-08085-8. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41146115/. Acesso em: 25 jul. 2026.

ZHENG, K.; SHEN, Z.; CHEN, Z.; CHE, C.; ZHU, H. Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education. **BMC Medical**

**Education**, [London], v. 24, n. 1, 1003, 13 set. 2024. DOI: https://doi.org/10.1186/s12909-024-05977-z. Disponível em: https://pubmed.ncbi.nlm.nih.gov/39272041/. Acesso em: 25 jul. 2026.

**Contatos:** marianaserra.duarte@mackenzista.com.br e carlos.fernandes@mackenzie.br