



INTELIGÊNCIA ARTIFICIAL NA GRADUAÇÃO MÉDICA: UMA REVISÃO DE ESCOPO DAS APLICAÇÕES EDUCACIONAIS

ARTIFICIAL INTELLIGENCE IN UNDERGRADUATE MEDICAL EDUCATION: A SCOPING REVIEW OF EDUCATIONAL APPLICATIONS

Flávia Silva e Oliveira^{1*}, Beatriz Pires Carcute², Emily Caroline Alves Martins², Júlia Castro Andrade², Luana Travassos Batista³

¹Doutora em Enfermagem pela Universidade Federal de Goiás (UFG), Goiânia (GO), Brasil; Docente da Faculdade de Medicina, Universidade Federal de Catalão (UFCAT), Catalão (GO), Brasil; ²Acadêmica do curso de Medicina, Faculdade de Medicina, Universidade Federal de Catalão (UFCAT), Catalão (GO), Brasil; ³Doutora em Imunologia Básica e Aplicada pela Universidade Federal do Amazonas (UFAM), Manaus (AM), Brasil; Professora substituta do Centro Universitário de Mineiros (UNIFIMES), Campus Trindade, Trindade (GO), Brasil.

***Autor correspondente:** Flávia Silva e Oliveira – **Email:** flavia04silva@gmail.com

Recebido: 11 ago. 2025

Aceito: 06 jan. 2026

Editores-chefes: Dr. Leonardo Pestillo de Oliveira e Dr. Mateus Dias Antunes

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution (<https://creativecommons.org/licenses/by/4.0/>), que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.



RESUMO: O objetivo foi mapear a aplicação da inteligência artificial (IA) na graduação médica, com ênfase nas ferramentas utilizadas, desfechos de aprendizagem, desempenho e lacunas da literatura. As buscas foram realizadas no PubMed, Scopus, Web of Science, EBSCOhost (CINAHL/MEDLINE), ScienceDirect e Google Scholar, incluindo estudos empíricos em inglês, espanhol e português, sem restrição temporal. A seleção dos estudos foi realizada em duplicata utilizando o Rayyan, e os achados foram descritos a partir do referencial TPACK. Dos 605 registros, 58 estudos (2018–2024) foram incluídos, majoritariamente de países de alta renda. Predominaram delineamentos quantitativos, com amostras de pequeno a médio porte. A IA foi aplicada principalmente ao ensino teórico, avaliação educacional e treinamento em simulações, frequentemente associada a melhorias no desempenho acadêmico e habilidades objetivas. Conclui-se que a IA demonstra potencial educacional, mas permanece limitada por heterogeneidade metodológica, lacunas teóricas e desigualdades contextuais, ressaltando a necessidade de integração curricular estruturada, ética e longitudinal.

PALAVRAS-CHAVE: Aprendizagem. Educação Médica. Ensino. Estudante. Inteligência Artificial.

ABSTRACT: The objective was to map the application of artificial intelligence (AI) in undergraduate medical education, focusing on the tools used, learning outcomes, performance, and gaps in the literature. Searches were conducted in PubMed, Scopus, Web of Science, EBSCOhost (CINAHL/MEDLINE), ScienceDirect, and Google Scholar, including empirical studies in English, Spanish, and Portuguese, without temporal restrictions. Study selection was performed in duplicate using Rayyan, and findings were described based on the TPACK framework. Of 605 records, 58 studies (2018–2024) were included, mostly from high-income countries. Quantitative designs predominated, with small to medium sample sizes. AI was primarily applied to theoretical instruction, educational assessment, and simulation-based training, often linked to improvements in academic performance and objective skills. It is concluded that AI demonstrates educational potential but remains limited by methodological heterogeneity, theoretical gaps, and contextual inequalities, highlighting the need for structured, ethical, and longitudinal curricular integration.

KEYWORDS: Artificial Intelligence. Learning. Medical Education. Student. Teaching.

1. INTRODUÇÃO

A inteligência artificial (IA) vem transformando rapidamente a área da saúde e a educação médica ao replicar funções cognitivas humanas e aumentar a eficiência dos processos.¹ Na área da saúde, a IA contribui para diagnósticos mais precisos e para a avaliação prognóstica por meio da análise de grandes volumes de dados², além de ter adquirido relevância crescente na promoção da saúde e na prevenção de doenças, que constituem prioridades centrais da saúde pública contemporânea. Iniciativas baseadas em IA têm sido amplamente utilizadas para estimular comportamentos saudáveis, mudanças no estilo de vida e ações preventivas, especialmente no contexto das doenças crônicas não transmissíveis.³

Na educação médica, a IA tem possibilitado ambientes de aprendizagem seguros, abrangentes e interativos, incluindo simulações, sistemas automatizados de feedback e ferramentas de avaliação adaptativa.³ Estudos internacionais relatam sua utilização na graduação médica em diferentes regiões do mundo, como América do Norte, Europa e Ásia, apoiando a preparação para exames, o treinamento de habilidades clínicas, o fornecimento de feedback e atividades curriculares diversas. Essas tecnologias apresentam potencial para democratizar o acesso a recursos educacionais, padronizar a formação e promover maior equidade nas oportunidades de aprendizagem.^{4,5} Quando integrada aos currículos de graduação, a IA pode contribuir para a preparação de futuros médicos voltados ao cuidado preventivo, às estratégias de saúde populacional e à prática clínica mediada por tecnologias digitais.^{6,7}

Apesar do interesse global crescente, as iniciativas que envolvem o uso da IA na educação médica variam amplamente entre países e instituições. A maioria dos estudos é oriunda de países de alta renda, havendo evidências limitadas provenientes de contextos de baixa e média renda, o que revela importantes desequilíbrios geográficos e contextuais. Ademais, as iniciativas identificadas variam desde projetos-piloto isolados e disciplinas eletivas até treinamentos baseados em simulação e avaliações educacionais apoiadas por IA.

Nesse cenário de rápida expansão, a adoção da IA na educação médica frequentemente ocorre de forma não coordenada e conceitualmente difusa, sendo impulsionada mais pela disponibilidade tecnológica do que por necessidades pedagógicas claramente definidas. Como consequência, observam-se implementações fragmentadas, com delineamentos instrucionais, objetivos de aprendizagem e estratégias de avaliação heterogêneos.⁸⁻¹¹ Essa variabilidade dificulta a comparação entre os estudos e limita a tradução dos achados em diretrizes curriculares coerentes, deixando educadores sem orientações claras sobre como alinhar a IA aos currículos, aos resultados de aprendizagem e às competências profissionais esperadas.^{8,12,13}

Muitos estudos têm comparado ferramentas de IA, particularmente modelos de linguagem de grande escala (LLM), como o ChatGPT, ao desempenho de estudantes de medicina, sobretudo em exames ou tarefas de perguntas e respostas.¹⁴ No entanto, essas comparações raramente são contextualizadas em estruturas curriculares ou pedagógicas mais amplas, o que limita sua relevância para a tomada de decisão educacional.^{8,13,19,20} Embora a IA possa atingir desempenho comparável ou até superior ao dos estudantes, os resultados são heterogêneos e sensíveis a fatores como formulação de prompts, idioma e formatos de avaliação, o que suscita preocupações quanto à confiabilidade, validade e significado educacional desses achados.¹⁵

Além disso, a evidência disponível tende a priorizar métricas isoladas de desempenho, com atenção limitada aos processos de aprendizagem, à integração pedagógica ou à coerência curricular.^{5,14,16-18} A expertise do corpo docente, a prontidão institucional e os contextos locais também são fatores críticos para a integração efetiva da IA, mas permanecem pouco explorados na literatura.^{8,12,21}

Nesse contexto, modelos conceituais fundamentados em teorias da aprendizagem tornam-se essenciais para esclarecer como a IA se articula com a pedagogia e com os conteúdos da formação médica. O referencial do *Technological Pedagogical Content Knowledge* (TPACK)²² oferece uma lente analítica pertinente ao enfatizar a relação integrada entre tecnologia, pedagogia e conhecimento disciplinar.

Dessa forma, esta revisão de escopo tem como objetivo mapear a aplicação da inteligência artificial na graduação médica, identificando as ferramentas utilizadas, os efeitos sobre a aprendizagem e o desempenho dos estudantes, bem como as lacunas existentes na literatura. Ao organizar as evidências sob uma perspectiva educacional, busca-se subsidiar o desenho curricular e a prática docente, apoiando educadores e instituições na navegação entre as oportunidades e os desafios da incorporação da IA em ambientes acadêmicos e clínicos cada vez mais digitalizados.

2. MÉTODOS

O protocolo desta revisão foi elaborado de acordo com as diretrizes PRISMA-P, refinado pela equipe de pesquisa e registrado na Open Science Framework (<https://doi.org/10.17605/OSF.IO/PZERT>). Esta revisão de escopo foi conduzida e relatada conforme as recomendações do PRISMA-ScR. Foi desenvolvida uma estratégia de busca abrangente adaptadas a cada base de dados. O processo de seleção dos estudos é apresentado no fluxograma PRISMA-ScR (Figura 1). Foram incluídos estudos publicados em inglês, espanhol e português, sem restrição temporal.

2.1 PERGUNTA DE PESQUISA

O framework População–Conceito–Contexto (PCC) foi aplicado para mapear sistematicamente as evidências sobre inteligência artificial (IA) na educação médica de graduação.

A questão norteadora: “Qual é o estado atual da aplicação da IA no processo de ensino–aprendizagem para docentes e estudantes de Medicina, e de que forma ela aprimora os desfechos educacionais?” foi operacionalizada da seguinte forma: a População incluiu docentes de Medicina e/ou estudantes de graduação em Medicina; o Conceito compreendeu tecnologias de IA (por exemplo, sistemas tutoriais adaptativos, plataformas automatizadas de avaliação e mecanismos inteligentes de feedback), sua integração curricular e funções pedagógicas; e o Contexto concentrou-se nos efeitos da IA sobre o desempenho discente, a carga de trabalho docente e o desenho curricular, bem como nos desafios e oportunidades técnicos, pedagógicos e éticos associados à sua implementação.

2.2 CRITÉRIOS DE ELEGIBILIDADE

Foram incluídos estudos quantitativos, qualitativos e de métodos mistos, revisados por pares, que apresentassem dados empíricos sobre aspectos educacionais relacionados à IA na graduação médica, tais como desfechos de aprendizagem, desenvolvimento de habilidades, integração curricular, métodos de ensino, avaliação ou engajamento discente. Foram excluídos estudos envolvendo exclusivamente pós-graduandos ou profissionais formados; protocolos de pesquisa; artigos retratados; e estudos que mencionassem IA sem analisar sua aplicação educacional. Fontes não acadêmicas também foram excluídas.

Revisões sistemáticas elegíveis foram incluídas, e suas listas de referências foram examinadas. Quando múltiplas publicações se referiam ao mesmo conjunto de dados, foi selecionado o estudo metodologicamente mais completo.

2.3 FONTES E ESTRATÉGIA DE BUSCA

Em 9 de dezembro de 2024, foram realizadas buscas nas bases de dados PubMed, Scopus, Web of Science, EBSCOhost (CINAHL/MEDLINE), ScienceDirect e Google Scholar. As estratégias de busca combinaram termos MeSH e palavras-chave relacionadas a docentes de Medicina, estudantes de Medicina, ensino-aprendizagem, inteligência artificial, machine learning e deep learning, com adaptações específicas para cada base de dados (Apêndice I).

A literatura cinzenta e os anais de eventos científicos foram recuperados por meio do Google Scholar, e as listas de referências de revisões relevantes foram examinadas manualmente. Todos os registros foram importados para o software Zotero, e os duplicados foram removidos.

2.4 SELEÇÃO DOS ESTUDOS

Antes do processo de triagem, a equipe, composta por quatro pesquisadores, realizou um estudo-piloto com 20 referências para alinhar os critérios de elegibilidade e os procedimentos de extração de dados, com reuniões de calibração após cada etapa. Dois revisores realizaram de forma independente a triagem dos títulos e resumos no software Rayyan, seguida da avaliação em duplicata dos textos completos elegíveis. Divergências foram resolvidas por consenso ou com a participação de um terceiro revisor.

2.5 EXTRAÇÃO E SÍNTESE DOS DADOS

Foi utilizado um formulário estruturado e previamente testado em estudo-piloto para a extração das informações-chave dos estudos, incluindo autoria, ano e país de publicação, delineamento do estudo, características metodológicas, ferramentas de IA empregadas, principais achados e relevância educacional. A extração dos dados foi realizada de forma independente por dois revisores, e eventuais discrepâncias foram resolvidas por meio de discussão com um terceiro revisor, assegurando consistência e acurácia.

Para subsidiar a interpretação educacional das evidências apresentadas na Tabela 1, foram definidos campos analíticos previamente estabelecidos. Os delineamentos dos estudos seguiram a terminologia reportada pelos próprios autores; quando não explicitados, a classificação baseou-se na abordagem metodológica predominante. Os principais achados corresponderam aos desfechos educacionais primários apresentados na seção de Resultados de cada estudo, identificados por meio de leitura integral dos textos, refletindo o principal efeito da intervenção com IA conforme definido pelos autores originais.

As ferramentas de IA foram identificadas a partir das descrições metodológicas e referem-se às tecnologias de inteligência artificial utilizadas em cada intervenção. Estas incluíram: LLMs (por exemplo, GPT-3.5, GPT-4, ChatGPT, Bard, Claude, Bing Chat, Copilot, PaLM-2 e Google Gemini); LLMs específicos ou personalizados (por exemplo, AnatomyGPT); modelos clássicos de machine learning (como support vector machines, k-nearest neighbours e motores de raciocínio baseado em casos); arquiteturas de deep learning (redes neurais convolucionais, recorrentes e em conjunto); sistemas híbridos combinando abordagens orientadas por dados e modelos mecanísticos (por exemplo, modelos ANN-PBPK); plataformas integradas de IA que fornecem feedback automatizado ou tutoria baseada em simulação (como Virtual Operative Assistant, ICEMS, Hepius e o indicador de qualidade GE EF); e conteúdos curriculares relacionados à IA que abordaram conceitos ou desenho de sistemas sem o uso operacional

direto de uma ferramenta específica (por exemplo, disciplinas teóricas, cursos sobre IA ou simulações de desenho de sistemas de apoio à decisão clínica baseados em IA).

Os estudos foram agrupados tematicamente em quatro categorias: (1) modelos de linguagem de grande escala; (2) sistemas automatizados de feedback; (3) plataformas de simulação e ferramentas de avaliação de exames/provas; e (4) módulos de integração curricular. Os resultados foram apresentados de forma descritiva e em tabelas, e foi realizada uma análise qualitativa para identificar tendências, padrões e lacunas na literatura. O referencial TPACK (*Technological Pedagogical Content Knowledge*)²² foi utilizado como lente analítica para categorizar implicitamente os estudos em três domínios: tecnológico (uso das ferramentas de IA), pedagógico (estratégias de ensino e avaliação) e de conteúdo (conhecimento clínico e médico), bem como em suas interseções, incluindo a aplicação pedagógica da tecnologia e sua integração com os conteúdos curriculares.

3. RESULTADOS

3.1 CARACTERÍSTICAS DOS ESTUDOS

Um total de 555 estudos foi identificado nas bases de dados selecionadas, complementado por 50 estudos provenientes da literatura cinzenta. A abrangência da estratégia de busca foi confirmada, uma vez que não foram identificados estudos adicionais a partir das listas de referências dos artigos incluídos. O software Zotero foi utilizado para a remoção de duplicatas, seguida de verificação no Rayyan.

Dos 536 títulos e resumos triados, 145 estudos foram selecionados para leitura na íntegra, resultando na inclusão de 58 estudos (Figura 1).

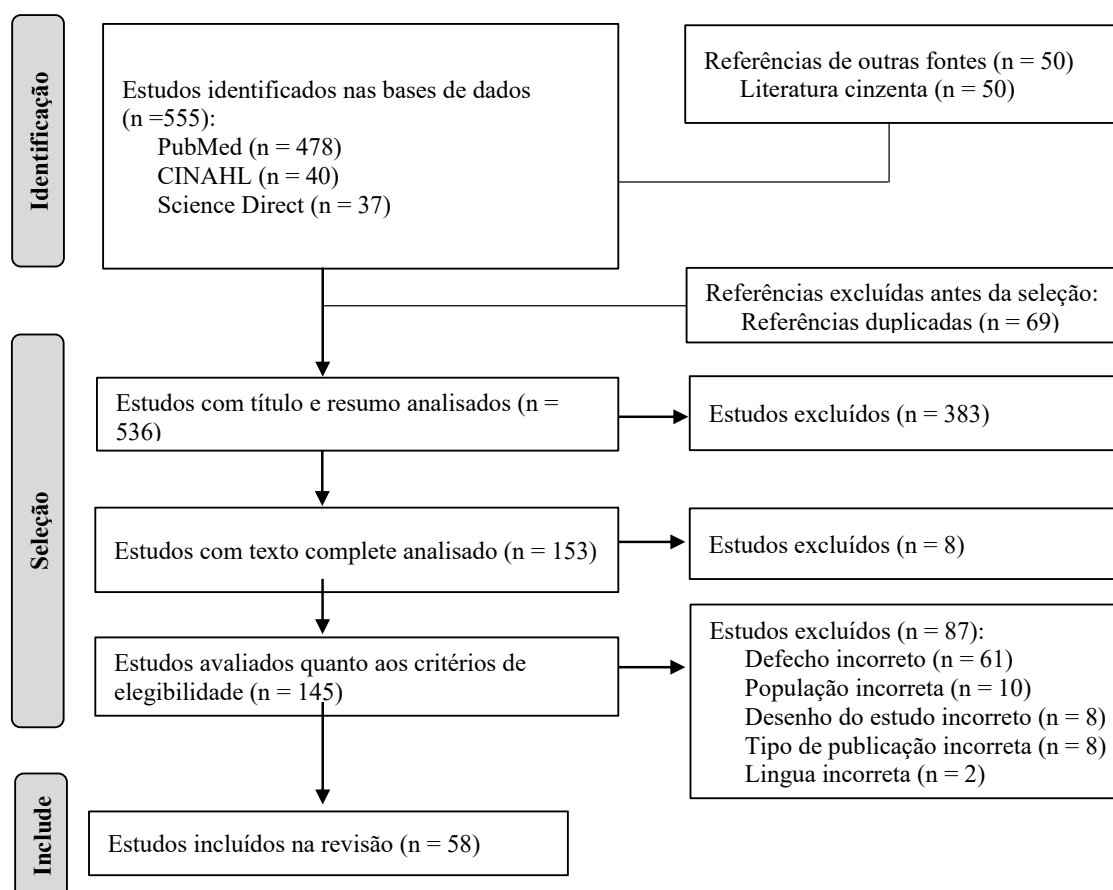


Figura 1. Diagrama de fluxo PRISMA-ScR do processo de seleção dos estudos.

A análise descritiva dos 58 estudos incluídos (Tabela 1) revelou que a maioria foi conduzida entre 2018 e 2024, com predominância na América do Norte, responsável por aproximadamente 33% (19/58) do total. Essa região incluiu estudos realizados nos Estados Unidos e Canadá.^{8,9,16,19,23-27}

A Europa surge em seguida, com aproximadamente 28% (16/58) dos estudos.^{5,10-13,17,20,38-44}

A Ásia correspondeu a cerca de 22% (13/58).^{18,21,45-52} A América Latina contribuiu com 5% (3/58), com estudos realizados no Brasil, Chile e México.⁵³⁻⁵⁵ O Oriente Médio representou aproximadamente 5% (3/58),⁵⁶⁻⁵⁸ enquanto a África correspondeu a 3% (2/58).^{59,60} Por fim, a Oceânia foi representada por 2% (1/58), com um estudo conduzido na Austrália.⁶¹

Quanto ao delineamento dos estudos, os estudos transversais foram os mais frequentes, representando 26% (15/58) do total.^{14,16,18,48,53,54,57-59,62} Em seguida, os delineamentos quase-experimentais corresponderam a 17% (10/58).^{9,25,26,32,43,45,51,52,55,63} Os estudos observacionais ^{5,12,17,20,27,30,31,42} e ensaios clínicos randomizados^{34,35,39,41,46,50,56,67} representaram cerca de 14% (8/58), respectivamente. Os estudos experimentais ^{19,23,28,29,33} e mistos^{10,11,17,27,63} totalizaram 9% (5/58), respectivamente. Os estudos descritivos representaram 7% (4/58)^{13,20,61,61}, enquanto o único estudo teórico-conceitual foi o de Houssaini et al., 2024 (2%; 1/58).⁶⁰

Em relação à abordagem metodológica, houve predominância de estudos quantitativos, presentes em 76% (44/58) dos casos.^{23,36,48,49,53,59,62,65} Os estudos de métodos mistos representaram 17% (10/58)^{10,11,17,27,63} e os estudos qualitativos corresponderam a 7% (4/58).^{21,49,60,64}

Os tamanhos amostrais variaram de 6 participantes³² a 500,^{19,44} com média de 132,4 participantes e mediana de 74,0. Metade dos estudos (50%) apresentou amostras pequenas a moderadas (menos de 100 participantes), enquanto 8 estudos (≈21%) incluíram mais de 200 participantes. Cerca de 20% (12/58) dos estudos não relataram o tamanho da amostra.

Observou-se predominância de estudantes de medicina do terceiro ano.^{26,41,59} Alguns estudos incluíram estudantes do primeiro ano⁴⁷, do quarto ano^{56,64} e, com menor frequência, do quinto ano.⁶⁶

Além dos graduandos, alguns estudos envolveram pacientes padronizados⁴¹, modelos de inteligência artificial, como chatbots⁶³, profissionais de saúde experientes³², equipes cirúrgicas²⁹ e residentes médicos.⁵¹

Quadro 1. Síntese das evidências sobre inteligência artificial na educação médica de graduação: características e achados dos estudos.

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
Meyer; Riese; Streichert, 2024 ⁵ / Alemanha	Estudo observacional comparativo	GPT-3.5 e GPT-4 responderam a 937 questões em texto de três exames de licenciamento alemães; os escores foram comparados aos das coortes reais de candidatos.	O GPT-4 obteve pontuação dentro dos percentis mais altos dos estudantes em todos os exames; o GPT-3.5 apresentou desempenho inferior, mas refletiu as taxas de aprovação dos alunos.	GPT-3.5, GPT-4	O GPT-4 pode atuar como parceiro de treino de alto nível, desde que as respostas sejam verificadas.
Zare et al., 2024 ^{14, 23} / Irã	Avaliação transversal	Três conjuntos de 200 questões de múltipla escolha (persa e inglês) foram submetidos ao GPT-3.5, GPT-4 e Bard, e comparados às notas anonimizadas de	O GPT-4 superou claramente os estudantes em persa e apresentou desempenho semelhante em inglês; o GPT-3.5	GPT-3.5, GPT-4, Bard	O GPT-4 favorece a prática avaliativa bilíngue; o idioma influencia o desempenho.

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
		436 estudantes iranianos.	falhou em todos os exames em persa.		
Abbas; Rehman; Rehman, 2024⁸ / Estados Unidos	Estudo computacional experimental	163 questões de múltipla escolha do NBME, de acesso público, foram inseridas individualmente em quatro modelos de linguagem de grande escala.	O GPT-4 respondeu corretamente 100% das questões; o segundo melhor (GPT-3.5) alcançou 82%.	GPT-4, GPT-3.5, Claude, Bard	O GPT-4 é o benchmark atual, mas ainda requer verificação humana para erros.
Abid et al., 2023²⁴ / Estados Unidos	Avaliação quase experimental de curso	Dezenove estudantes do último ano concluíram um eletivo autoguiado de IA com duração de um mês; foram analisados questionários de confiança pré e pós-curso e produtos finais dos projetos.	A confiança autorreferida em conceitos de IA/ML aumentou 66%; um quarto dos estudantes continuou desenvolvendo projetos de IA após o curso.	Curso teórico baseado em IA	Cursos eletivos curtos, baseados em projetos, podem ampliar a literacia em IA.
Burke et al., 2024²⁵ / Estados Unidos	Estudo observacional retrospectivo	ChatGPT-3.5 e pacientes padronizados pontuaram independentemente 168 notas clínicas de estudantes do primeiro ano, usando uma rubrica de 85 itens.	A taxa de erro do ChatGPT foi de 1%, em comparação a 7% dos avaliadores humanos.	GPT-3.5	A IA pode fornecer feedback rápido e consistente sobre documentação em texto livre.
Cianciolo; Lavoie; Parker, 2022²⁶ / Estados Unidos	Estudo observacional retrospectivo	700 redações de justificativa diagnóstica foram reavaliadas por humanos; um modelo de PLN foi treinado com métricas textuais para prever essas pontuações.	As pontuações do modelo correlacionaram-se fortemente com as avaliações humanas ($r \approx 0,7$) e com outros indicadores acadêmicos.	PLN / ML customizado	A correção automatizada de redações sobre raciocínio clínico é viável e válida.
Collins; Black; Rarey, 2024¹⁹ / Estados Unidos	Estudo computacional experimental	Um "AnatomyGPT" customizado, baseado em textos de código aberto, foi testado contra o ChatGPT em 60 questões no estilo NBME.	A acurácia foi semelhante, mas o AnatomyGPT forneceu mais citações de fontes em suas explicações.	AnatomyGPT (GPT-4)	GPTs adaptados ao domínio podem atuar como tutores ricos em referências.
Fazlollahi et al., 2023³⁷ / Canadá	Estudo de coorte retrospectivo	Estudantes realizaram cinco ressecções tumorais em RV; um tutor de IA (VOA) selecionou quatro alvos de feedback e monitorou 54 variáveis de movimento.	As habilidades-alvo melhoraram, mas algumas métricas não-alvo apresentaram deriva, evidenciando efeitos não intencionais da aprendizagem.	Virtual Operative Assistant	É necessária supervisão especializada quando a IA orienta a prática cirúrgica.
Fazlollahi et al., 2022³⁶ / Canadá	Ensaio clínico randomizado	Setenta estudantes foram alocados para feedback por IA (VOA), feedback remoto de especialista ou ausência de feedback durante cinco tarefas	O grupo VOA alcançou o melhor desempenho objetivo sem aumento da carga cognitiva.	Virtual Operative Assistant	A tutoria por IA baseada em métricas pode superar especialistas em ambientes de simulação.

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
		neurocirúrgicas em RV.			
Gin et al., 2022²⁷ / Estados Unidos	Estudo de métodos mistos	Um modelo LSTM bidirecional foi treinado com 3.328 feedbacks narrativos de internato para prever níveis de confiabilidade; técnicas de IA destacaram características linguísticas.	O modelo associou padrões específicos de feedback a níveis mais altos ou mais baixos de confiabilidade.	Rede neural LSTM	A IA pode analisar grandes volumes de feedback para orientar docentes quanto a linguagem utilizada na supervisão dos estudantes.
Krive et al., 2023⁸ / Estados Unidos	Estudo de coorte prospectivo	Vinte estudantes do último ano cursaram um curso eletivo online de 4 semanas integrando casos de IA; quizzes, projetos e reflexões foram analisados.	A média dos quizzes foi de 97%; os estudantes relataram transição da teoria para a aplicação prática da IA.	Curso sobre IA	Cursos eletivos bem estruturados podem incorporar o pensamento aplicado em IA aos currículos.
Ledwos et al., 2022²⁸ / Canadá	Estudo experimental	Cinquenta cirurgiões e estudantes repetiram uma ressecção tumoral subpial em RV cinco vezes; métricas selecionadas por IA acompanharam curvas de aprendizagem.	Cinco de seis métricas mostraram progressão clara e distinguiram níveis de expertise.	Seleção métrica por k-NN	Métricas objetivas baseadas em inteligência artificial sustentam o acompanhamento adaptativo da progressão do aprendizado.
Mirchi et al., 2020²⁹ / Canadá	Estudo experimental de validação	Cinquenta participantes realizaram uma tarefa em RV; um VOA baseado em SVM classificou desempenho de especialista vs. iniciante e gerou painéis de feedback.	Acurácia do classificador de 92%, sensibilidade de 100% e especificidade de 82%.	Support Vector Machine	Feedback explicável por IA favorece confiança e autorreparo de habilidades.
Morjaria et al., 2023³⁰ / Canadá	Estudo retrospectivo não controlado	ChatGPT-3.5 respondeu a 30 questões discursivas curtas estratificadas por Bloom; seis tutores avaliaram as respostas em relação a mínimos estudantis e médias históricas.	O ChatGPT superou estudantes limítrofes, mas ficou abaixo da média histórica da turma.	GPT-3.5	Questões discursivas sem supervisão apresentam risco de validade quando LLMs estão acessíveis.
Nagaraj et al., 2022³¹ / Estados Unidos	Estudo observacional retrospectivo	229 vídeos de estudantes realizando nós cirúrgicos foram codificados como aprovado/reprovado; CNNs 2D e 3D detectaram erros de empunhadura e nós.	Acurácia ~90%, porém com menor sensibilidade, resultando na perda de alguns maus desempenhos.	Redes neurais convolucionais	A tutoria automatizada por vídeo é promissora, mas deve reduzir falsos negativos.
Park et al., 2024⁹ / Estados Unidos	Piloto quase experimental	Dez estudantes do primeiro ano cursaram um curso eletivo paralelo de IA, sem créditos, integrado a quatro blocos pré-clínicos; questionários de confiança por bloco.	A confiança em temas de IA aumentou significativamente em todos os blocos; satisfação geral de 4,3/5.	ML clássico + demonstrações CNN	Mesmo pilotos pequenos mostram como a IA pode ser integrada às ciências básicas.

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
Ruberto et al., 2021³² / Canadá	Estudo experimental de prova de conceito	Dois médicos e dois estudantes realizaram simulação de ressuscitação em RA por 10 minutos; uma DNN multitarefa usou ECG e GSR para ajustar a gravidade em tempo real.	O sistema modulou com sucesso a dificuldade do cenário e foi avaliado como envolvente.	Rede neural profunda multitarefa	A RA adaptativa pode ajustar níveis de estresse para otimizar a aprendizagem.
Scherr et al., 2023³³ / Estados Unidos	Estudo de caso descritivo	ChatGPT-3.5, com engenharia de prompts, produziu três simulações interativas de ACLS/UTI; os autores refinaram iterativamente realismo e acurácia.	O ChatGPT gerou cenários dinâmicos e ramificados, responsivos às entradas do aprendiz.	GPT-3.5	Simulações clínicas gratuitas e sob demanda são viáveis.
Sridharan; Sequeira, 2024¹⁶ / Reino do Bahrein	Auditoria transversal de ferramentas	ChatGPT-3.5, Sage e Claude criaram objetivos de aprendizagem e itens avaliativos em farmacologia da hipertensão.	Os objetivos alinharam-se consistentemente à Taxonomia de Bloom; a qualidade dos itens variou, exigindo edição docente.	GPT-3.5, Sage, Claude	LLMs podem redigir conteúdos curriculares, mas requerem revisão especializada.
Yilmaz et al., 2023³⁴ / Canadá	Ensaio clínico randomizado	120 estudantes, em quatro grupos, praticaram ressecção bimanual em RV com feedback numérico, visual, visuoespacial ou nenhum.	Os grupos com feedback visual e visuoespacial atingiram mais critérios de especialista após cinco repetições.	Motor métrico baseado em SVM	Feedback visual por IA acelera habilidades cirúrgicas mais que números isolados.
Yilmaz et al., 2024³⁵ / Canadá	Ensaio clínico randomizado duplo-cego	97 estudantes receberam feedback em tempo real por IA (ICEMS), feedback de especialista em tempo real ou apenas pós-hoc durante cinco ressecções em RV.	O grupo ICEMS apresentou maiores escores compostos e menor carga cognitiva.	Intelligent Continuous Expertise Monitoring System	O feedback em tempo real por IA pode superar a instrução tradicional por especialistas.
Bajcetic et al., 2024³⁸ / Sérvia	Revisão narrativa metodológica	Discussão de simulações baseadas em ANN e PBPK usando dados reais de pacientes para modelar ADME/PD de fármacos.	Modelos aprimorados por IA produziram previsões PK/PD de alta-fidelidade, reduzindo testes em animais.	Híbridos ANN + PBPK	Simulações farmacológicas com pacientes virtuais enriquecem o ensino de farmacologia.
Birks et al., 2024¹⁰ / Reino Unido	Estudo de métodos mistos pré-pós	68 estudantes aprenderam engenharia de prompts com o Google Gemini para elaborar questões SBA; questionário de confiança e grupo focal.	A confiança no uso de IA para elaboração de questões aumentou significativamente; houve valorização da ferramenta e preocupação com acurácia.	Google Gemini	Oficinas de “avaliação invertida” desenvolvem alfabetização em IA e escrita de prompts.
Borchert et al., 2023²⁰ / Reino Unido	Estudo descritivo avaliativo	ChatGPT resolveu todo o banco de prática do SJT do Reino Unido de 2023; respostas pontuadas com rubrica oficial e mapeadas aos	Escore global de 76%, equilibrado entre domínios de profissionalismo; melhor desempenho em tarefas de	GPT-3.5	O ChatGPT pode apoiar a revisão de ética e profissionalismo, mas requer validação humana.

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
		domínios do GMC.	ranqueamento.		
Coskun et al., 2024³⁹ / Turquia	Ensaio clínico randomizado duplo-cego	74 estudantes compararam vinhetas geradas pelo ChatGPT com vinhetas humanas e responderam a 15 MCQs do ChatGPT.	As vinhetas foram avaliadas de forma equivalente; 11/15 MCQs da IA atenderam a métricas aceitáveis.	GPT-3.5	O ChatGPT pode elaborar casos clínicos utilizáveis, economizando tempo docente.
Driesnack et al., 2024¹² / Alemanha	Estudo observacional prospectivo	46 voluntários participaram de curso curto gamificado sobre uso racional de antibióticos, apoiado pelo GPT-4; testes pré, pós e em 12 meses.	As taxas de acerto aumentaram 74–83% e se mantiveram por um ano; 92% consideraram o feedback da IA útil.	GPT-4	Gamificação associada ao GPT promove aprendizagem duradoura e confiança.
Friederichs et al., 2023¹⁷ / Alemanha	Avaliação observacional	ChatGPT respondeu a 400 MCQs do Progress Test; resultados comparados às normas alemãs dos anos 1–6.	65,5% de acertos, superior à maioria dos anos 1–3 e semelhante a estudantes avançados.	GPT-3.5	LLMs são adequados para autoavaliação factual, com necessidade de enquadramento crítico.
Holderried et al., 2024⁴⁰ / Alemanha	Estudo prospectivo unicêntrico	Chatbot GPT-4 simulou pacientes em 106 entrevistas escritas de 8 minutos e gerou feedback estruturado; comparação com avaliador humano.	99% das respostas foram clinicamente plausíveis; concordância do feedback $\kappa = 0,83$.	GPT-4	O GPT-4 pode oferecer treino estruturado de anamnese em larga escala.
Lippitsch et al., 2024⁴¹ / Alemanha	Ensaio clínico randomizado cruzado controlado	168 estudantes alternaram entre ViPATalk (chatbot avatar baseado em regras) e role-play entre pares em três casos.	O ViPATalk foi não inferior ao role-play quanto à completude e comunicação.	Chatbot baseado em regras AIML	Chatbots de pacientes virtuais podem substituir com segurança sessões de role-play.
Moglia et al., 2022⁴² / Itália	Estudo observacional preditivo	176 iniciantes realizaram cinco tarefas de cirurgia robótica em RV; DNNs em ensemble predizeram tempo e tentativas até proficiência, sendo comparadas aos modelos de Floresta Aleatória (RF) e Árvores de Regressão com Gradiente Boosting (GBRT).	As DNNs identificaram trainees com dificuldade mais cedo e com maior precisão que outros modelos.	Redes neurais profundas em ensemble	A predição precoce por IA permite personalizar recursos de tutoria.
Riedel et al., 2023⁴³ / Alemanha	Estudo computacional quase experimental	ChatGPT respondeu a dois conjuntos de MCQs em ginecologia/obstetrícia; especialistas avaliaram acurácia e densidade de insights.	Acurácia semelhante à dos estudantes; explicações continham muitos insights não óbvios.	GPT-3.5	O ChatGPT agrega riqueza explicativa que favorece o autoestudo.
Soulage et al., 2024¹³ / França	Estudo quantitativo comparativo	GPT-3.5 realizou um exame de fisiologia de nível de mestrado junto a 24 estudantes; correção cega.	Mediana da IA de 75% vs. 59% dos humanos; a IA foi aprovada em 100% das vezes.	GPT-3.5	Os formatos avaliativos devem evoluir frente às capacidades dos LLMs.
Thomae et al., 2024¹¹ /	Estudo exploratório	ChatGPT integrado a três unidades de	Os estudantes consideraram os	GPT-3.5	Integrações simples baseadas em prompts

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
Suíça	de métodos mistos	ensino sobre placebo/nocebo; questionários de percepção pré e pós-intervenção.	textos claros e confiáveis; o curso foi avaliado como útil.		desenvolvem pensamento crítico sobre IA.
Tomova et al., 2024⁴⁴ / Alemanha	Estudo experimental computacional	GPT-4 e Bing Chat geraram feedback por tópico para 18 MCQs do Progress Test; seis especialistas avaliaram relevância (1–5).	Ambos os modelos obtiveram média >4/5; não foram detectadas alucinações.	GPT-4, Bing Chat	LLMs podem produzir feedback formativo de alta qualidade em larga escala.
Barui et al., 2022⁴⁵ / Índia	Estudo observacional comparativo	CNNs VGG16/Inception treinadas com 1.000 imagens histológicas e testadas em 300 imagens; 25 estudantes realizaram o mesmo teste.	Acurácia da CNN de 89% vs. 78% dos estudantes; a IA foi mais rápida e consistente.	CNN (VGG16, Inception)	A IA de reconhecimento de imagens fortalece sessões de histologia.
Cheng et al., 2020⁴⁶ / Taiwan	Ensaio clínico randomizado	36 estudantes do quinto ano receberam ensino de radiografia de fratura de quadril com ou sem IA; testes pré e pós.	O grupo IA melhorou de 76% para 89%; o controle não apresentou mudança significativa.	DenseNet-121 “HipGuide”	Tutores por IA com mapas de calor aumentam a acurácia interpretativa.
Ghosh et al., 2023⁴⁷ / Índia	Estudo quase experimental em sala de aula	ChatGPT-3.5 respondeu a 79 itens de bioquímica; respostas corrigidas junto à coorte de 100 estudantes do primeiro ano.	A IA ficou em 5º lugar na turma, mas foi mais lenta em MCQs de raciocínio.	GPT-3.5	LLMs podem atuar como “estudante adicional” para verificação imediata.
Huang et al., 2024⁴⁸ / Taiwan	Estudo transversal de desempenho	GPT-4 resolveu 600 MCQs bilíngues do exame de licenciamento (10 disciplinas básicas).	Acurácia global de 87,5%, superando o corte de 60% em todas as disciplinas.	GPT-4	Ferramenta viável de autoestudo bilíngue para ciências básicas.
Huh, 2023¹⁸ / Coreia	Comparação transversal	GPT-3.5 e 77 calouros responderam a 79 MCQs de parasitologia; um especialista avaliou as explicações.	A IA obteve 61% vs. 91% dos estudantes; explicações corretas foram aceitáveis.	GPT-3.5	O GPT-3.5 tem desempenho inferior em domínios específicos, mas modela boas explicações.
Kinoshita et al., 2024⁴⁹ / Japão	Estudo observacional retrospectivo	U-Net treinado com 740 imagens cirúrgicas rotuladas destacou nervos; vídeo anotado de 1 min vs. aula expositiva.	Estudantes com o vídeo de IA dobraram os escores de reconhecimento neural.	U-Net	Vídeos curtos anotados por IA melhoram o reconhecimento anatômico.
Li et al., 2023⁵⁰ / China	Ensaio clínico randomizado + entrevistas	30 estudantes do quinto ano aprenderam morfologia sanguínea com ou sem plataforma de mapas de calor por IA.	Acurácia do grupo IA aumentou de 76% para 89%; estudantes valorizaram pistas visuais.	CNN DeepCyto	Mapas de calor guiados por IA encurtam curvas de aprendizagem visual.
Su et al., 2024⁵¹ / Taiwan	Validação quase experimental	54 estudantes usaram o sistema adaptativo MEKAS; grupo controle com estudo convencional.	O grupo experimental apresentou ganho significativo de conhecimento e alta usabilidade.	Motor CBR baseado em conhecimento	Testes adaptativos por competência personalizam trajetórias de estudo.
Zheng et	Estudo quase	66 estudantes	O grupo IA obteve	GPT-3.5,	Cenários com IA

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
al., 2024⁵² / China	experimental em sala de aula	receberam ensino cardiológico baseado em cenários convencionais ou com IA (ChatGPT + Midjourney).	escores superiores em teoria, Mini-CEX e pensamento crítico.	Midjourney	aumentam engajamento e habilidades clínicas.
Al-Worafi et al., 2023²¹ / Malásia	Estudo qualitativo de análise de conteúdo	Docentes consultaram o ChatGPT sobre desenho de módulo de farmacoterapia em doenças infecciosas.	Avaliações variaram de 50–92%; útil para estruturação, porém superficial em alguns pontos.	ChatGPT	Utilizar o ChatGPT como gerador de ideias, não como currículo pronto.
Alessi et al., 2024⁵³ / Brasil	Estudo transversal de desempenho	GPT-3.5 respondeu a 333 MCQs do Teste de Progresso (2021–2023); comparação com médias nacionais.	Média da IA de 68% vs. 51% dos estudantes; melhor desempenho em Saúde Coletiva.	GPT-3.5	O ChatGPT supera todos os anos de estudantes brasileiros em testes factuais.
Alonso-Silveiro et al., 2018⁵⁴ / México	Estudo transversal de desenvolvimento	Treinador laparoscópico de baixo custo com Raspberry-Pi rastreou movimentos dos instrumentos; ANN classificou expertise dos usuários ao longo de cinco dias.	Acurácia de classificação de 91%; curvas claras de aprendizagem para iniciantes.	Rede neural artificial	Treinadores de baixo custo baseados em IA permitem o acompanhamento objetivo da aquisição de habilidades psicomotoras.
Castellano et al., 2024⁵⁵ / Chile	Estudo quase experimental de curso	131 calouros usaram app móvel gamificado de anatomia com recomendador neural e assistente virtual NLP.	A ferramenta melhorou desempenho e engajamento; estudantes valorizaram torneios e dicas personalizadas.	Recomendador NN sequencial, NLP Dialogflow	Gamificação com IA fortalece aprendizagem autorregulada em anatomia.
Aronovitz et al., 2024⁵⁶ / Israel	Ensaio clínico randomizado paralelo	60 estudantes realizaram teste de aquisição de imagens POCUS com ou sem indicador de qualidade por IA em tempo real.	O grupo IA obteve escores maiores ($p = 0,002$) e maior chance de vistas apicais corretas.	Indicador de qualidade EF (GE)	Feedback em tempo real por IA melhora significativamente a qualidade de imagens.
Meo et al., 2024⁵⁷ / Arábia Saudita	Estudo diagnóstico transversal	Google Bard respondeu a 60 questões OSPE baseadas em imagens de seis órgãos.	Acurácia global de 57%, variando de 28% (rim) a 90% (osso).	Google Bard	O Bard apresenta habilidade moderada em identificação visual; requer refinamento.
Mavrych; Ganguly; Bolgova, 2024⁵⁸ / Arábia Saudita	Comparação transversal	Seis LLMs responderam a 50 MCQs de anatomia de membro superior e geraram cenários; avaliação por dois docentes.	GPT-4 obteve maior escore (61%); Bard/Gemini tiveram os menores (33%).	GPT-4, GPT-3.5, Copilot, PaLM-2, Bard, Gemini	LLMs avançados auxiliam na elaboração de itens, mas não substituem docentes.
Cherif et al., 2024⁵⁹ / Tunísia	Estudo transversal de desempenho	ChatGPT-3.5 respondeu a 45 MCQs de pneumologia sob dois formatos de prompt; comparação com 244 estudantes.	Prompt contextualizado aprovou no exame (51%) e superou 62% dos estudantes.	GPT-3.5	Engenharia de prompts influencia fortemente o sucesso em exames.
Houssaini et al., 2024⁶⁰ / Marrocos	Estudo conceitual de design	Desenvolvimento do framework DRIIPT (Design Thinking + IA) para curso de inovação no primeiro ano.	Relatos de aumento de criatividade, pensamento crítico e fluência em GenAI.	ChatGPT	A integração de design thinking e GenAI promove habilidades de inovação.
Xie et al.,	Estudo	ChatGPT-4, Bard e	O ChatGPT obteve	GPT-4, Bard,	O ChatGPT atua como

Autores (Ano – País)	Desenho do estudo	Métodos	Principais achados	Ferramenta(s) de IA	Relevância educacional
2024⁶¹ / Austrália	descritivo com avaliação de especialistas	Bing AI responderam a cenários cirúrgicos escalonados; avaliados por legibilidade e qualidade.	os maiores escores de legibilidade, qualidade e acurácia.	Bing AI	adjunto confiável para autoaprendizagem clínica inicial.
Gan et al., 2024⁶² / China	ECR duplo-cego de desenho paralelo	129 estudantes estudaram ortopedia com ou sem auxílio do ChatGPT por duas semanas; acompanhamento de testes e provas.	O grupo ChatGPT teve melhor desempenho imediato e em avaliações subsequentes.	GPT-4	O ChatGPT melhora desempenho avaliativo de curto e médio prazo.
Baglivo et al., 2023⁶³ / Itália	Estudo de viabilidade em sala de aula	36 estudantes compararam respostas de cinco chatbots sobre vacinação em sessão de 2h; discussão e satisfação.	Chatbots superaram estudantes e estimularam debates críticos.	ChatGPT, Bing Chat, Chatsonic, Bard, YouChat	Uso guiado de chatbots aumenta engajamento e pensamento crítico.
Tsopra et al., 2022⁶⁴ / França	Estudo descritivo de desenvolvimento curricular	Quinze estudantes projetaram protótipos de sistemas de decisão clínica por IA em curso eletivo.	Satisfação de 4,4/5; vários seguiram projetos em IA.	Design de CDSS simulado	Colocar estudantes como designers aprofunda letramento em saúde digital.
Bharatha et al., 2024⁶⁵ / Barbados	Estudo transversal de desempenho	GPT-4 e estudantes responderam a 304 MCQs classificados por nível de Bloom.	GPT-4 obteve 74% vs. 67% dos estudantes; melhor em conhecimento factual.	GPT-4	LLMs são potentes para recordação, mas mais fracos em níveis avaliativos.
Furlan et al., 2021⁶⁶ / Itália	Estudo pré- e pós-experimental	Quinze graduandos usaram simulador de paciente virtual Hepius com ITS; MCQs pré e pós.	Escore totais e centrais aumentaram significativamente ($p<0,001$).	SLSTM NLP + ITS	VPS móvel pode melhorar rapidamente o raciocínio diagnóstico.

ACLS – Suporte Avançado de Vida em Cardiologia; ADME – Absorção, Distribuição, Metabolismo e Excreção; AI – Inteligência Artificial; AIML – Linguagem de Marcação para Inteligência Artificial; ANN – Rede Neural Artificial; AR – Realidade Aumentada; Bard – Chatbot Google Bard/Gemini; CBR – Raciocínio Baseado em Casos; CDSS – Sistema de Apoio à Decisão Clínica; CNN – Rede Neural Convolutiva; DNN – Rede Neural Profunda; DRIIPT – Framework DRIIPT (Design Thinking, Alinhamento Construtivo e IA Generativa); EF – Fração de Ejeção; ENT – Otorrinolaringologia; GBRT – Árvores de Regressão com Gradiente Boosting; GMC – General Medical Council (Conselho Médico Geral do Reino Unido); GSR – Resposta Galvânica da Pele; GPT-3.5 – Transformador Generativo Pré-treinado, versão 3.5; GPT-4 – Transformador Generativo Pré-treinado, versão 4; ICEMS – Sistema Inteligente de Monitoramento Contínuo de Expertise; ICU – Unidade de Terapia Intensiva; ITS – Sistema de Tutoria Inteligente; k-NN – k-Vizinhos Mais Próximos; LLM – Modelo de Linguagem de Grande Escala; LSTM – Memória de Longo e Curto Prazo; MCQ – Questão de Múltipla Escolha; Mini-CEX – Mini-Exercício Clínico Avaliativo; ML – Aprendizado de Máquina; NBME – National Board of Medical Examiners; NLP – Processamento de Linguagem Natural; OB/GYN – Obstetrícia e Ginecologia; OSPE – Exame Prático Estruturado Objetivo; PBPK – Farmacocinética Baseada em Fisiologia; PD – Farmacodinâmica; POCUS – Ultrassonografia à Beira do Leito; RF – Floresta Aleatória; RCT – Ensaio Clínico Randomizado; SJT – Teste de Julgamento Situacional; SAQ – Questão de Resposta Curta; Seq. NN – Rede Neural Sequencial; SLSTM – LSTM Siamês; SVM – Máquina de Vetores de Suporte; U-Net – Rede Neural Convolutiva em “U”; VPS – Simulador de Paciente Virtual; VOA – Assistente Operatório Virtual; VR – Realidade Virtual.

3.2 CATEGORIZAÇÃO TEMÁTICA DOS ESTUDOS

3.2.1 MODELOS DE LINGUAGEM DE GRANDE ESCALA

Entre os 58 estudos incluídos, 20 (34,5%) avaliaram LLMs em contextos de exames formais, comparando as pontuações e taxas de aprovação do ChatGPT-3.5, ChatGPT-4, Google Bard e outros LLMs com o desempenho de estudantes de Medicina.^{5,14,48,59}

Em exames médicos padronizados, o GPT-4 demonstrou desempenho superior, com acurácia variando de 70,6% a 100%, superando de forma consistente o ChatGPT-3.5 e outros modelos, como Bard e Claude.^{23,62} O desempenho dos estudantes apresentou ampla variação, com médias entre 50,8% e 90,8%, dependendo do tipo de exame e do contexto linguístico.^{18,53} Em contraste, o ChatGPT-3.5 alcançou taxas de acurácia entre 51,1%, abaixo do limiar de aprovação de 60% em um exame de pneumologia sem orientação contextual, e 82,2% em avaliações no estilo NBME, além de apresentar desempenho inferior à média dos estudantes em disciplinas como parasitologia (60,8% vs. 90,8%).^{18,23,59}

3.2.2 SISTEMAS AUTOMATIZADOS DE FEEDBACK

Dezesseis estudos (27,6%) abordaram sistemas automatizados de feedback para habilidades clínicas e cirúrgicas, avaliando escores compostos (por exemplo, ICEMS e OSATS), métricas de segmentação (Dice, IoU) e avaliações da qualidade de imagem em POCUS.^{34,40,49,56}

Sistemas de feedback clínico e cirúrgico, como o Virtual Operative Assistant (VOA) e o Intelligent Continuous Expertise Monitoring System (ICEMS), resultaram em redução de erros e melhora de desempenho entre 10% e 20%, com significância estatística ($P < 0,01$).^{34,67} Outros estudos também destacaram o uso da IA para feedback automatizado, incluindo simulações baseadas no GPT-4⁴⁰, correção de anotações clínicas²⁵, feedback formativo no treinamento cirúrgico²⁹ e análise baseada em IA de feedback narrativo²⁷.

3.2.3 PLATAFORMAS DE SIMULAÇÃO E FERRAMENTAS DE AVALIAÇÃO PARA EXAMES

Nove estudos (15,5%) descreveram plataformas de simulação e tutoria virtual, empregando testes pré e pós-intervenção, medidas de adaptação da carga cognitiva e classificadores de aprendizado de máquina para avaliar a efetividade pedagógica.^{29,32,51,54,66}

Plataformas orientadas por IA, como Hepius (simulador de paciente virtual com ITS, demonstrando ganho médio de 15 pontos percentuais nos escores de raciocínio clínico)⁶⁶; HipGuide (mapas de calor em radiografias de quadril, aumentando a acurácia de 75,7% para 88,9%)⁴⁶; ICEMS (feedback cirúrgico em tempo real em simulação de ressecção tumoral, com melhora de 12 pontos)³⁴; Virtual Operative Assistant (VOA) (tutor baseado em SVM para simulação neurocirúrgica, alcançando ganhos de até 15 pontos sem aumento da carga cognitiva)³⁶; EF Quality Indicator (ferramenta de feedback em POCUS que tornou 2,3 vezes mais provável a aquisição correta de imagens apicais)⁵⁶; DeepCyto, sistema de IA para morfologia celular (uma CNN para classificação de células sanguíneas, elevando a acurácia em 12 pontos)⁵⁰; e U-Net (modelo de segmentação de vídeos cirúrgicos que dobrou os escores de reconhecimento anatômico)⁴⁹, demonstraram ganhos entre 12 e 20 pontos percentuais ($P < 0,01$).

Essas aplicações voltadas à avaliação evidenciaram maior consistência, escalabilidade e eficiência na mensuração do desempenho discente, ao mesmo tempo em que suscitaram considerações relevantes sobre validade, confiabilidade e o risco de dependência excessiva de sistemas automatizados de avaliação.³⁴

3.2.4 MÓDULOS DE INTEGRAÇÃO CURRICULAR

Onze estudos (19,0%) examinaram a integração de ferramentas de IA em currículos e cursos híbridos, utilizando questionários validados para avaliar satisfação discente, progresso percebido na aprendizagem e autoconfiança em conceitos relacionados à IA.^{8,9,11,24} Dois estudos adicionais (3,4%)

empregaram abordagens qualitativas, incluindo análise de conteúdo, para explorar percepções sobre benefícios, riscos e fatores que influenciam a adoção da IA na educação médica.^{21,64}

A integração de módulos de IA nos currículos, por meio de disciplinas eletivas,^{8,24} oficinas práticas focadas na geração de questões com LLMs¹⁰ e sessões de journal club que combinaram discussão de artigos com demonstrações de programação⁹, esteve associada a um aumento de 50% a 66% na confiança dos estudantes no uso de ferramentas de IA, além de alto desempenho acadêmico, com médias $\geq 90\%$ em questionários e avaliações de projetos.^{8,24}

Achados qualitativos provenientes de docentes e instrutores indicaram que 80% a 90% perceberam as soluções de IA como úteis para o planejamento curricular e o desenvolvimento de materiais educacionais, ao mesmo tempo em que enfatizaram a necessidade de supervisão crítica para evitar resultados imprecisos ou superficiais. Em um estudo de análise de conteúdo, 85% dos educadores recomendaram o ChatGPT como ferramenta de apoio, desde que objetivos de aprendizagem e itens de avaliação fossem cuidadosamente revisados.²¹ De modo semelhante, 80% dos estudantes e instrutores participantes de um curso de design de sistemas de apoio à decisão clínica baseados em IA (IA-CDSS) recomendaram a atividade, ressaltando a importância da supervisão humana contínua.⁶⁴

3.3 FRAMEWORK TPACK

A análise desses estudos revela uma convergência prática com os domínios do modelo TPACK (Fig. 2), embora nenhum deles tenha referenciado explicitamente a sua aplicação. As intervenções baseadas em IA refletem conhecimento tecnológico (por exemplo, uso de LLMs, simuladores e ferramentas de feedback), conhecimento pedagógico (por exemplo, feedback em tempo real, aprendizagem personalizada e simulações clínicas) e conhecimento do conteúdo (por exemplo, ensino de anatomia, diagnóstico e habilidades cirúrgicas). Os estudos que integraram os três domínios, como oficinas que utilizaram LLMs para abordar conteúdos médicos ou tutores inteligentes que forneceram feedback clínico, ilustram o potencial do TPACK como estrutura orientadora para o desenvolvimento futuro de currículos aprimorados por IA na educação médica.

3.4 APLICAÇÕES NA EDUCAÇÃO MÉDICA

As evidências sugerem que as ferramentas de IA podem ser empregadas para: (1) apoiar a aprendizagem autodirigida e a preparação para exames, por meio do reforço da aquisição de conhecimentos;^{5,53,59} (2) fornecer feedback imediato e padronizado no treinamento de habilidades clínicas e cirúrgicas;^{34,37,56} (3) aprimorar simulações interativas e o ensino de técnicas de imagem;^{33,49,57} (4) automatizar a correção de avaliações objetivas e personalizar trajetórias de aprendizagem;^{14,19,44} e (5) auxiliar no desenvolvimento de competências críticas em IA e no planejamento curricular.^{8,11,21,39} Contudo, é necessária supervisão pedagógica contínua para assegurar a validade educacional e o cumprimento de padrões éticos.^{13,30}

4. DISCUSSÃO

Dos 58 estudos analisados (2018–2024), 83% tiveram origem na América do Norte, Europa e Ásia, enquanto apenas 15% foram provenientes da América Latina, Oriente Médio, África e Oceania, evidenciando desigualdades globais significativas na produção científica, provavelmente relacionadas a diferenças no acesso tecnológico, financiamento e prioridades institucionais.

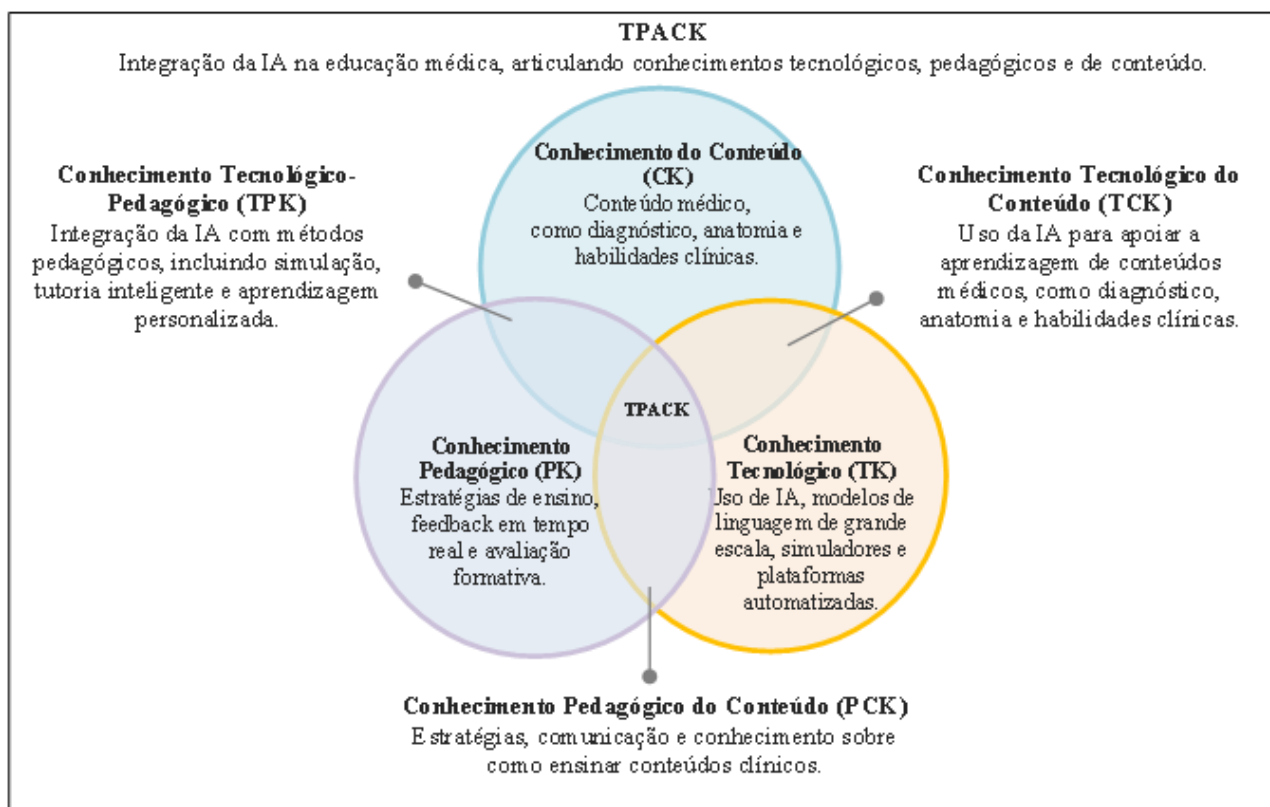


Figura 2. Modelo conceitual TPACK para a aplicação da inteligência artificial na educação médica.

Fonte: os autores.

A maioria dos estudos adotou delineamentos transversais ou quase-experimentais e concentrou-se em ferramentas de IA, como modelos de linguagem de grande escala, sistemas automatizados de feedback e módulos curriculares orientados por IA. No entanto, muitas investigações apresentaram limitações relacionadas a amostras pequenas, desenhos de centro único e curta duração, restringindo a validade externa e a avaliação de desfechos de longo prazo.^{8,66} Observa-se ainda uma lacuna relevante de estudos longitudinais conduzidos em ambientes clínicos reais, de currículos fundamentados teoricamente e de avaliações sobre retenção de conhecimento, competências interprofissionais e dimensões éticas da integração da IA.

Os estudantes relataram consistentemente alta percepção da relevância da IA para a prática médica (94,8–99,1%)⁶⁸⁻⁷²; entretanto, o ensino formal em IA permanece limitado, resultando em familiaridade insuficiente com a terminologia técnica (50–60%) e preparo inadequado para aplicação clínica.^{69,70,72} Apesar dessa lacuna, o interesse discente pela educação em IA é elevado, com apoio a currículos obrigatórios⁷⁰, preferência por conteúdos que abordem mecanismos da IA, ética e riscos⁷¹ e defesa de oficinas interdisciplinares e práticas.⁷⁰ Persistem preocupações relacionadas à responsabilidade legal e às limitações da IA em cenários clínicos imprevisíveis, especialmente quando podem ocorrer desfechos adversos.⁶⁸

Uma limitação crítica refere-se ao desempenho reduzido dos LLMs na geração de conteúdos precisos e sensíveis ao contexto em idiomas distintos do inglês, levantando preocupações quanto à adaptação transcultural e ao uso adequado de terminologia técnica quando os dados de treinamento são predominantemente anglófonos.⁵⁹ Na ausência de estratégias de mitigação direcionadas, os benefícios educacionais da IA podem permanecer concentrados em regiões de língua inglesa, ampliando a marginalização de contextos de baixa e média renda e linguisticamente diversos. Em contrapartida, a IA tem demonstrado potencial para automatizar materiais de educação em saúde, possibilitando a

produção rápida de recursos personalizados que podem melhorar a adesão ao tratamento e os desfechos em saúde.¹¹

Para além da eficiência educacional, a integração da IA na graduação em Medicina possui implicações relevantes para a promoção da saúde e a prevenção de doenças. Ferramentas apoiadas por IA podem fortalecer competências relacionadas a aconselhamento preventivo, estratificação de risco, educação do paciente e comunicação baseada em evidências, centrais às estratégias de saúde populacional.^{3,6,7} Por meio de trajetórias de aprendizagem personalizadas, identificação precoce de lacunas de conhecimento e treinamento baseado em simulação com foco em decisões preventivas, a IA pode indiretamente favorecer melhores comportamentos em saúde, adesão e equidade na atenção.^{3,11,73} Esses benefícios, contudo, dependem de currículos eticamente fundamentados e de forte supervisão docente para evitar amplificação de vieses, desinformação e dependência tecnológica, especialmente entre populações vulneráveis.^{5,59}

Em avaliações teóricas, os LLMs apresentaram desempenho satisfatório em provas de múltipla escolha e testes de progresso, apoiando seu uso como ferramentas de autoestudo e revisão.^{16,41,48} Persistem, entretanto, limitações como respostas inconsistentes²⁰, viés algorítmico⁵, desempenho inferior em raciocínio clínico complexo⁴⁷, incapacidade de expressar incerteza¹⁷ e justificativas insuficientes das respostas.¹⁸ Embora melhorias algorítmicas recentes¹⁴ e avanços tecnológicos⁶⁵ tenham mitigado parte desses problemas, o uso acrítico da IA, sem o desenvolvimento paralelo de competências reflexivas e críticas, pode comprometer o raciocínio clínico.¹⁷

Além dos testes teóricos, a IA tem sido aplicada de forma eficaz em simulações clínicas e no treinamento autodirigido de habilidades, ao fornecer feedback contextualizado em tempo real e acelerar curvas de aprendizagem, particularmente nas fases iniciais da formação.^{12,40,46,56} Essas ferramentas também favorecem a identificação precoce de estudantes com dificuldades, melhoram o desempenho acadêmico e reduzem a carga cognitiva.^{34,42}

Docentes percebem a IA como um recurso pedagógico valioso, capaz de aumentar o engajamento discente¹¹, automatizar a geração de conteúdos²¹, reduzir a carga de avaliação³¹ e oferecer feedback frequente e direcionado. A IA também viabiliza vinhetas clínicas realistas e estratégias instrucionais otimizadas.^{16,39,44} A implementação bem-sucedida, contudo, requer capacitação docente, tempo para revisão curricular e consenso institucional sobre o papel da IA, assegurando que ela complemente, e não substitua, a interação humana e as dimensões emocionais do ensino.^{8,12} Alcançar esse equilíbrio exige colaboração entre especialistas em IA, educadores médicos e clínicos, para enfrentar desafios técnicos, éticos e culturais.^{13,20,63,65}

Profissionais de saúde antecipam que especialidades como radiologia, dermatologia e patologia liderarão a adoção da IA, seguidas por cirurgia e atenção primária na próxima década.⁶⁸ Em contextos clínicos, a IA tem demonstrado acurácia prognóstica e preditiva terapêutica comparável ou superior à de especialistas humanos, reduzindo subjetividade e variabilidade interobservador.⁷⁴ Modelos preditivos têm sido aplicados à fertilização in vitro, desfechos neonatais, parto prematuro, peso fetal e risco de aneuploidias⁷⁵, enquanto sistemas de apoio à decisão clínica baseados em IA já estão consolidados em diagnóstico por imagem, oncologia, predição de resposta terapêutica e gestão de dados clínicos.⁷⁶ No âmbito da saúde populacional, a IA auxilia na previsão de hospitalizações evitáveis, lacunas de rastreamento e na automação de triagem, revisão de prontuários, fluxos administrativos e aconselhamento de pacientes.⁷³

De modo geral, esses achados evidenciam a necessidade urgente de currículos estruturados, multicêntricos e interdisciplinares que abordem não apenas competências técnicas em IA, mas também aspectos éticos, legais e de letramento digital, aliados à educação médica continuada. Sem tais

adaptações, instituições acadêmicas correm o risco de obsolescência diante das rápidas transformações na educação e na prática médica.^{13,76}

Esta revisão é limitada pela ausência de avaliação formal da qualidade metodológica, pela impossibilidade de quantificar tamanhos de efeito e pela natureza rapidamente evolutiva do campo. Ademais, a predominância de estudos pequenos e de centro único limita a validade externa e a generalização para contextos educacionais e sistemas de saúde diversos.

4.1 IMPLICAÇÕES PRÁTICAS

Programas de graduação em Medicina devem estabelecer políticas institucionais claras que regulamentem o uso educacional de modelos de linguagem de grande escala, especialmente em contextos avaliativos, a fim de proteger a integridade acadêmica e assegurar a verificação adequada de conteúdos gerados por IA.^{5,14,30} Quando devidamente reguladas, essas ferramentas também podem apoiar competências relevantes à promoção da saúde, como aconselhamento preventivo e comunicação em saúde baseada em evidências, reduzindo riscos associados à desinformação e às iniquidades.^{3,77}

Sistemas de feedback e ferramentas de simulação baseados em IA podem ser integrados como adjuvantes ao treinamento de habilidades clínicas e cirúrgicas, oferecendo feedback padronizado e oportuno e monitoramento objetivo do desempenho. Para além da aquisição técnica, essas aplicações podem contribuir para a promoção da saúde e a prevenção de doenças, ao fortalecer o reconhecimento precoce de riscos e a tomada de decisão preventiva, desde que a supervisão docente assegure alinhamento com princípios centrados no ser humano e mitigue efeitos de aprendizagem não intencionais.^{6,37,56,67}

Por fim, a integração eficaz da IA requer modelos curriculares longitudinais, baseados em frameworks, e não disciplinas eletivas isoladas. Tais abordagens devem incorporar letramento em IA, formação ética e desenvolvimento docente, com adaptação aos contextos locais e linguísticos, de modo a promover implementação equitativa e evitar o reforço de desigualdades em saúde observadas nas iniciativas atuais de promoção da saúde orientadas por IA.^{3,7}

5. CONCLUSÃO

Esta revisão de escopo mapeou 58 estudos (2018–2024) sobre inteligência artificial na educação médica de graduação e organizou as evidências em quatro categorias temáticas: modelos de linguagem de grande escala, utilizados principalmente para preparação para exames, resolução de questões e geração de conteúdo; sistemas automatizados de feedback, voltados ao desenvolvimento de habilidades clínicas e cirúrgicas; plataformas de simulação e ferramentas orientadas à avaliação, incluindo tutoria virtual e avaliação assistida por IA; e módulos de integração curricular, como disciplinas eletivas e oficinas focadas em letramento e aplicação da IA.

De modo geral, as intervenções apoiadas por IA estiveram mais consistentemente associadas à melhora do desempenho em avaliações baseadas em conhecimento e, em contextos de simulação, a ganhos em métricas objetivas de desempenho, sustentados por feedback em tempo real. Os módulos de integração curricular frequentemente relataram aumento da confiança dos estudantes e da percepção de preparo em relação aos conceitos e ao uso da IA. Diversos estudos também enfatizaram que produtos gerados por IA requerem revisão especializada para assegurar a validade educacional, especialmente quando utilizados para feedback ou avaliação.

5.1 RECOMENDAÇÕES PARA ESTUDOS FUTUROS

Com base nas lacunas identificadas nesta revisão de escopo, pesquisas futuras devem adotar delineamentos multicêntricos e multilíngues, abrangendo diferentes especialidades médicas e níveis de formação, bem como comparar distintas versões de LLMs e estratégias de prompt para avaliar seu impacto educacional. A incorporação de dados multimodais e a análise do uso autônomo de prompts pelos estudantes podem otimizar a interação humano-IA.^{5,20,48,59}

Para suprir a evidência limitada sobre desenvolvimento de competências, são necessários estudos longitudinais em ambientes clínicos reais, utilizando métricas validadas para comparar métodos de feedback baseados em IA e tradicionais^{34,56}, além de avaliar raciocínio clínico, habilidades interprofissionais e retenção de conhecimento em longo prazo.^{34,56}

Por fim, a ausência de modelos curriculares orientados por teoria destaca a necessidade de integração da IA ancorada em teorias consolidadas da aprendizagem, como a ACT-R e a andragogia, a fim de orientar tanto o design instrucional quanto a gestão de riscos.^{19,40} A integração ética também requer a avaliação de vieses algorítmicos, efeitos comportamentais e riscos como dependência cognitiva e redução do pensamento crítico.^{37,52}

5.2 FINANCIAMENTO

Este estudo não recebeu financiamento de fontes públicas, privadas ou institucionais.

REFERÊNCIAS

1. Dave M, Patel N. Artificial intelligence in healthcare and education. *British Dental Journal*. 2023;234(10): 761–764. <https://doi.org/10.1038/s41415-023-5845-2>.
2. Jaworski A, Jasiński D, Jaworski W, Hop A, Janek A, Sławińska B, et al. Comparison of the Performance of Artificial Intelligence Versus Medical Professionals in the Polish Final Medical Examination. *Cureus*. 2024;16(8): e66011. <https://doi.org/10.7759/cureus.66011>.
3. Yousefi F, Naye F, Ouellet S, Yameogo AR, Sasseville M, Bergeron F, et al. Artificial Intelligence in Health Promotion and Disease Reduction: Rapid Review. *Journal of Medical Internet Research*. 2025;27: e70381–e70381. <https://doi.org/10.2196/70381>.
4. Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, bin Saleh K, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*. 2023;19(8): 1236–1242. <https://doi.org/10.1016/j.sapharm.2023.05.016>.
5. Meyer A, Riese J, Streichert T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Medical Education*. 2024;10(1). <https://doi.org/10.2196/50965>.
6. Smith A, Arena R, Bacon SL, Faghy MA, Grazi G, Raisi A, et al. Recommendations on the use of artificial intelligence in health promotion. *Progress in Cardiovascular Diseases*. 2024;87: 37–43. <https://doi.org/10.1016/j.pcad.2024.10.003>.
7. Qian Y, Siau KL. Advances in IoT, AI, and Sensor-Based Technologies for Disease Treatment, Health Promotion, Successful Ageing, and Ageing Well. *Sensors*. 2025;25(19): 6207. <https://doi.org/10.3390/s25196207>.
8. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: artificial intelligence in medical education. *JAMIA Open*. 2023;6(2). <https://doi.org/10.1093/jamiaopen/ooad037>.

9. Park SH, Pinto-Powell R, Thesen T, Lindqwister A, Levy J, Chacko R, et al. Preparing healthcare leaders of the digital age with an integrative artificial intelligence curriculum: a pilot study. *Medical Education Online*. 2024;29(1). <https://doi.org/10.1080/10872981.2024.2315684>.
10. Birks S, Gray J, Darling-Pomranz C. Using artificial intelligence to provide a 'flipped assessment' approach to medical education learning opportunities. *Medical Teacher*. 2024;47(8): 1377. <https://doi.org/10.1080/0142159X.2024.2434101>.
11. Thomae A V., Witt CM, Barth J. Integration of ChatGPT Into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications. *JMIR Medical Education*. 2024;10. <https://doi.org/10.2196/50545>.
12. Driesnack S, Rücker F, Dietze-Jergus N, Bondarenko A, Pletz MW, Viehweger A. A practice-based approach to teaching antimicrobial therapy using artificial intelligence and gamified learning. *JAC-Antimicrobial Resistance*. 2024;6(4). <https://doi.org/10.1093/jacamr/dlae099>.
13. Soulage CO, Van Coppenolle F, GUEBRE-EGZIABHER F. The conversational AI 'ChatGPT' outperforms medical students on a physiology university examination. *Advances in Physiology Education*. 2024;48(4): 677–684. <https://doi.org/10.1152/advan.00181.2023>.
14. Zare S, Vafaeian S, Amini M, Farhadi K, Vali M, Golestani A. Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams. *Scientific Reports*. 2024;14(1): 28456. <https://doi.org/10.1038/s41598-024-79335-w>.
15. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Academic Medicine*. 2024;99(1): 22–27. <https://doi.org/10.1097/ACM.0000000000005439>.
16. Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. *BMC Medical Education*. 2024;24(1):431. <https://doi.org/10.1186/s12909-024-05365-7>.
17. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? *Medical Education Online*. 2023;28(1). <https://doi.org/10.1080/10872981.2023.2220920>.
18. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *Journal of Educational Evaluation for Health Professions*. 2023;20: 1. <https://doi.org/10.3352/jeehp.2023.20.1>.
19. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: A customized artificial intelligence application for anatomical sciences education. *Clinical Anatomy*. 2024;37(6): 661–669. <https://doi.org/10.1002/ca.24178>.
20. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the Situational Judgement Test—A Professional Dilemmas–Based Examination for Doctors in the United Kingdom. *JMIR Medical Education*. 2023;9. <https://doi.org/10.2196/48978>.
21. Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC. The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives. *JMIR Medical Education*. 2024;10(1). <https://doi.org/10.2196/47339>.
22. Mishra P, Koehler MJ. Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record: The Voice of Scholarship in Education*. 2006;108(6): 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>.
23. Abbas A, Rehman MS, Rehman SS. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. *Cureus*. 2024;16(3). <https://doi.org/10.7759/cureus.55991>.
24. Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J. AI Education for Fourth-Year Medical Students: Two-Year Experience of a Web-Based, Self-Guided Curriculum and Mixed Methods Study. *JMIR Medical Education*. 2024;10(1). <https://doi.org/10.2196/46500>.
25. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the Ability of a Large Language Model to Score Free-Text Medical Student Clinical Notes: Quantitative Study. *JMIR Medical Education*. 2024;10. <https://doi.org/10.2196/56342>.
26. Cianciolo AT, Lavoie N, Parker J. Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence. *Academic Medicine*. 2021;96(7): 1026–1035. <https://doi.org/10.1097/ACM.0000000000004010>.

27. Gin BC, ten Cate O, O'Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. *Medical Education*. 2022;56(3): 303–311. <https://doi.org/10.1111/medu.14696>.
28. Ledwos N, Mirchi N, Yilmaz R, Winkler-Schwartz A, Sawni A, Fazlollahi AM, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *Journal of Neurosurgery*. 2022;137(4): 1160–1171. <https://doi.org/10.3171/2021.12.JNS211563>.
29. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE*. 2020;15(2). <https://doi.org/10.1371/journal.pone.0229596>.
30. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an Undergraduate Medical Program. *Journal of Medical Education and Curricular Development*. 2023;10. <https://doi.org/10.1177/23821205231204178>.
31. Nagaraj MB, Namazi B, Sankaranarayanan G, Scott DJ. Developing artificial intelligence models for medical student suturing and knot-tying video-based assessment and coaching. *Surgical Endoscopy*. 2023;37(1): 402–411. <https://doi.org/10.1007/s00464-022-09509-y>.
32. Ruberto AJ, Rodenburg D, Ross K, Sarkar P, Hungler PC, Etemad A, et al. The future of simulation-based medical education: Adaptive simulation utilizing a deep multitask neural network. *AEM Education and Training*. 2021;5(3). <https://doi.org/10.1002/aet2.10605>.
33. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT Interactive Medical Simulations for Early Clinical Education: Case Study. *JMIR Medical Education*. 2023;9(1). <https://doi.org/10.2196/49877>.
34. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, Wang A, Makhani HH, Alsayegh A, et al. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems–A Four-Arm Randomized Control Trial. *Journal of Surgical Education*. 2023;81(2): 275–287. <https://doi.org/10.6084/m9.figshare.15132507.v1>.
35. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, Wang A, Makhani HH, Alsayegh A, et al. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems–A Four-Arm Randomized Control Trial. *Journal of Surgical Education*. 2024;81(2): 275–287. <https://doi.org/https://doi.org/10.1016/j.jsurg.2023.11.001>.
36. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students a randomized clinical trial. *JAMA Network Open*. 2022;5(2). <https://doi.org/10.1001/jamanetworkopen.2021.49008>.
37. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in Surgical Curriculum Design and Unintended Outcomes for Technical Competencies in Simulation Training. *JAMA network open*. 2023;6(9): e2334658. <https://doi.org/10.1001/jamanetworkopen.2023.34658>.
38. Bajčetić M, Mirčić A, Rakočević J, Đoković D, Milutinović K, Zaletel I. Comparing the performance of artificial intelligence learning models to medical students in solving histology and embryology multiple choice questions. *Annals of Anatomy - Anatomischer Anzeiger*. 2024;254: 152261. <https://doi.org/https://doi.org/10.1016/j.aanat.2024.152261>.
39. Coşkun Ö, Kiyak YS, Budakoğlu İİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher*. 2024;(2): 268–274. <https://doi.org/10.1080/0142159X.2024.2327477>.
40. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A Language Model-Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR Medical Education*. 2024;10. <https://doi.org/10.2196/59213>.
41. Lippitsch A, Steglich J, Ludwig C, Kellner J, Hempel L, Stoevesandt D, et al. Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars. *Computer Methods and Programs in Biomedicine*. 2024;244: 107964. <https://doi.org/https://doi.org/10.1016/j.cmpb.2023.107964>.
42. Moglia A, Morelli L, D'Ischia R, Fatucchi LM, Pucci V, Berchiolli R, et al. Ensemble deep learning for the prediction of proficiency at a virtual simulator for robot-assisted surgery. *Surgical Endoscopy*. 2022;36(9): 6473–6479. <https://doi.org/10.1007/s00464-021-08999-6>.

43. Riedel M, Kaefinger K, Stuehrenberg A, Ritter V, Amann N, Graf A, et al. ChatGPT's performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. *Frontiers in Medicine*. 2023;10. <https://doi.org/10.3389/fmed.2023.1296615>.
44. Tomova M, Roselló Atanet I, Sehy V, Sieg M, März M, Mäder P. Leveraging large language models to construct feedback from medical multiple-choice Questions. *Scientific Reports*. 2024;14(1). <https://doi.org/10.1038/s41598-024-79245-x>.
45. Barui S, Sanyal P, Rajmohan KS, Malik A, Dudani S. Perception without preconception: comparison between the human and machine learner in recognition of tissues from histological sections. *Scientific Reports*. 2022;12(1). <https://doi.org/10.1038/s41598-022-20012-1>.
46. Cheng CT, Chen CC, Fu CY, Chaou CH, Wu YT, Hsu CP, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights into Imaging*. 2020;11(1). <https://doi.org/10.1186/s13244-020-00932-0>.
47. Ghosh A, Maini Jindal N, Gupta VK, Bansal E, Kaur Bajwa N, Sett A. Is ChatGPT's Knowledge and Interpretative Ability Comparable to First Professional MBBS (Bachelor of Medicine, Bachelor of Surgery) Students of India in Taking a Medical Biochemistry Examination? *Cureus*. 2023;15(10). <https://doi.org/10.7759/cureus.47329>.
48. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. *Digital Health*. 2024;10. <https://doi.org/10.1177/20552076241233144>.
49. Kinoshita K, Maruyama T, Kobayashi N, Imanishi S, Maruyama M, Ohira G, et al. An artificial intelligence-based nerve recognition model is useful as surgical support technology and as an educational tool in laparoscopic and robot-assisted rectal cancer surgery. *Surgical Endoscopy*. 2024;38(9): 5394–5404. <https://doi.org/10.1007/s00464-024-10939-z>.
50. Li J, Ouyang J, Liu J, Zhang F, Wang Z, Guo X, et al. Artificial Intelligence-based online platform assists blood cell morphology learning: A mixed-methods sequential explanatory designed research. *Medical Teacher*. 2023;45(6): 596–603. <https://doi.org/10.1080/0142159X.2023.2190483>.
51. Su JM, Hsu SY, Fang TY, Wang PC. Developing and validating a knowledge-based AI assessment system for learning clinical core medical knowledge in otolaryngology. *Computers in Biology and Medicine*. 2024;178. <https://doi.org/10.1016/j.compbiomed.2024.108765>.
52. 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*. 2024;24(1). <https://doi.org/10.1186/s12909-024-05977-z>.
53. Alessi MR, Gomes HA, Lopes de Castro M, Terumy Okamoto C. Performance of ChatGPT in Solving Questions From the Progress Test (Brazilian National Medical Exam): A Potential Artificial Intelligence Tool in Medical Practice. *Cureus*. 2024;16(7). <https://doi.org/10.7759/cureus.64924>.
54. Alonso-Silverio GA, Pérez-Escamirosa F, Bruno-Sanchez R, Ortiz-Simon JL, Muñoz-Guerrero R, Minor-Martinez A, et al. Development of a Laparoscopic Box Trainer Based on Open Source Hardware and Artificial Intelligence for Objective Assessment of Surgical Psychomotor Skills. *Surgical Innovation*. 2018;25(4): 380–388. <https://doi.org/10.1177/1553350618777045>.
55. Castellano MS, Contreras-McKay I, Neyem A, Farfán E, Inzunza O, Ottone NE, et al. Empowering human anatomy education through gamification and artificial intelligence: An innovative approach to knowledge appropriation. *Clinical Anatomy*. 2024;37(1): 12–24. <https://doi.org/10.1002/ca.24074>.
56. Aronovitz N, Hazan I, Jedwab R, Shitrit I Ben, Quinn A, Wacht O, et al. The effect of real-time EF automatic tool on cardiac ultrasound performance among medical students. *PLoS ONE*. 2024;19(3). <https://doi.org/10.1371/journal.pone.0299461>.
57. Meo SA, AbuKhalaf AA, Meo MZS, Meo MOS, Ayub R, ElToukhy RA, et al. Role of artificial intelligence (Google bard) in morphological, histopathological, and radiological image identifications: Objective Structured Practical Examination (OSPE) type-based performance. *Saudi medical journal*. 2024;45(5): 531–536. <https://doi.org/10.15537/smj.2024.45.5.20240141>.
58. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis. *Clinical Anatomy*. 2024;38(2): 200–210. <https://doi.org/10.1002/ca.24244>.

59. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's Aptitude for Medical Education: Comparative Analysis With Third-Year Medical Students in a Pulmonology Examination. *JMIR Medical Education*. 2024;10: e52818. <https://doi.org/10.2196/52818>.
60. Squalli Houssaini M, Aboutajeddine A, Toughrai I, Ibrahimi A. Development of a design course for medical curriculum: Using design thinking as an instructional design method empowered by constructive alignment and generative AI. *Thinking Skills and Creativity*. 2024;52: 101491. <https://doi.org/https://doi.org/10.1016/j.tsc.2024.101491>.
61. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. *ANZ Journal of Surgery*. 2024;94(1–2): 68–77. <https://doi.org/10.1111/ans.18666>.
62. Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, et al. Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial. *Journal of Medical Internet Research*. 2024;26. <https://doi.org/10.2196/57037>.
63. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the Possible Use of AI Chatbots in Public Health Education: Feasibility Study. *JMIR Medical Education*. 2023;9. <https://doi.org/10.2196/51421>.
64. Tsopra R, Peiffer-Smadja N, Charlier C, Campeotto F, Lemogne C, Ruszniewski P, et al. Putting undergraduate medical students in AI-CDSS designers' shoes: An innovative teaching method to develop digital health critical thinking. *International Journal of Medical Informatics*. 2023;171: 104980. <https://doi.org/https://doi.org/10.1016/j.ijmedinf.2022.104980>.
65. Bharatha A, Ojeh N, Rabbi AMF, Campbell MH, Krishnamurthy K, Layne-Yarde RNA, et al. Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. *Advances in Medical Education and Practice*. 2024;15: 393–400. <https://doi.org/10.2147/AMEP.S457408>.
66. Furlan R, Gatti M, Menè R, Shiffer D, Marchiori C, Gaj Levra A, et al. A Natural Language Processing-Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study. *JMIR Medical Informatics*. 2021;9(4). <https://doi.org/https://doi.org/10.2196/24073>.
67. Fazlollahi AM, Bakhaider M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Network Open*. 2022;5(2). <https://doi.org/10.1001/jamanetworkopen.2021.49008>.
68. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? *Digital Health*. 2023;9. <https://doi.org/10.1177/20552076231152167>.
69. Buabbas AJ, Miskin B, Alnaqi AA, Ayed AK, Shehab AA, Syed-Abdul S, et al. Investigating Students' Perceptions towards Artificial Intelligence in Medical Education. *Healthcare*. 2023;11(9): 1298. <https://doi.org/10.3390/healthcare11091298>.
70. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: A global mixed-methods study of medical students' perspectives. *DIGITAL HEALTH*. 2022;8: 205520762210890. <https://doi.org/10.1177/20552076221089099>.
71. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical Students' Attitudes Toward AI in Medicine and their Expectations for Medical Education. *Journal of Medical Education and Curricular Development*. 2023;10. <https://doi.org/10.1177/23821205231219346>.
72. Al Saad MM, Shehadeh A, Alanazi S, Alenezi M, Abu alez A, Eid H, et al. Medical Students' Knowledge and Attitude Towards Artificial Intelligence: An Online Survey. *The Open Public Health Journal*. 2022;15. <https://doi.org/https://doi.org/10.2174/18749445-v15-e2203290>.
73. Lin S. A Clinician's Guide to Artificial Intelligence (AI): Why and How Primary Care Should Lead the Health Care AI Revolution. *The Journal of the American Board of Family Medicine*. 2022;35(1): 175–184. <https://doi.org/10.3122/jabfm.2022.01.210226>.
74. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*. 2019;16(11): 703–715. <https://doi.org/10.1038/s41571-019-0252-y>.

75. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorgiou AT, Jouannic JM. Contributions of Artificial Intelligence Reported in Obstetrics and Gynecology Journals: Systematic Review. *Journal of Medical Internet Research*. 2022;24(4): e35465. <https://doi.org/10.2196/35465>.
76. Rösler W, Altenbuchinger M, Baeßler B, Beissbarth T, Beutel G, Bock R, et al. An overview and a roadmap for artificial intelligence in hematology and oncology. *Journal of Cancer Research and Clinical Oncology*. 2023;149(10): 7997–8006. <https://doi.org/10.1007/s00432-023-04667-5>.
77. Uddin J, Feng C, Xu J. Health Communication on the Internet: Promoting Public Health and Exploring Disparities in the Generative AI Era. *Journal of Medical Internet Research*. 2025;27: e66032. <https://doi.org/10.2196/66032>.