RMMG
Revista Médica de Minas Gerais

ARTIGO DE REVISÃO

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

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

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

RESUMO

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

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

[1] Universidade de Pernambuco (UPE), Recife, Pernambuco, Brasil

Editor Associado Responsável:
José Maria Peixoto
Faculdade de Medicina da UNIFENAS. Belo Horizonte/MG, Brasil.

Autor correspondente:
Davi Ricardo Soares Gama de Amorim
Departamento de Medicina da UPE, Pernambuco.
Email: davi.amorim@upe.br

Fontes apoiadoras:
Não há.

Conflito de Interesses:
Não há.

Recebido em: 10 Setembro 2025.
Aprovado em: 12 Fevereiro 2026.
Data de Publicação: 09 Junho 2026.

DOI: 10.5935/2238-3182.2026e36204

ABSTRACT

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

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

Introdução

A Inteligência Artificial (IA) configura um sistema computacional que simula e sistematiza aspectos específicos da inteligência humana ou do comportamento inteligente, como aprendizagem, intelectualidade e resolução de problemas[1]. Esse termo foi empregado pela primeira vez em 1956 por John McCarthy na Conferência de Dartmouth. Segundo ele, a inteligência artificial era "a ciência e a engenharia de fabricar máquinas inteligentes"[2].

Nesse sentido, entende-se que a IA é composta por três paradigmas principais: o simbólico (baseado na lógica e no conhecimento), o estatístico (métodos probabilísticos e aprendizado de máquina) e o subsimbólico (inteligência embarcada e pesquisa). Esses paradigmas abordam vários domínios de problemas, como percepção, raciocínio, conhecimento, planejamento e comunicação[3]. Embora existam diferentes categorias de inteligência artificial, *Machine Learning* (ML)[4] destaca-se, que é um sistema que treina um modelo preditivo — identificando padrões em dados de entrada — e o utiliza para fazer previsões úteis a partir de dados novos e inéditos[1].

Porém, é importante ressaltar que nem todo algoritmo de aprendizado de máquina é universal e capaz de resolver todos os tipos de problemas. Assim, torna-se crucial selecionar conjuntos de dados apropriados para resolver um problema específico. Nesse sentido, as Redes Neurais Artificiais desempenham um papel significativo. São modelos matemáticos inspirados em estruturas neurais

biológicas, adquirindo capacidade computacional por meio do aprendizado[5].

Outra tarefa da inteligência artificial envolve o processamento de linguagem natural, que abrange a capacidade da máquina de compreender a linguagem humana. Porém, não consiste simplesmente em reconhecer as letras que formam uma palavra e depois combiná-las com uma definição; o algoritmo é usado para obter uma compreensão da sintaxe e da semântica para aproximar o significado de frases, sentenças ou parágrafos[4].

Devido à sua complexidade, embora a inteligência artificial esteja muitas vezes associada apenas a computadores ou robôs, a sua aplicação pode ser observada em vários domínios, abrangendo áreas como filosofia, psicologia, linguística, estatística[4] e até mesmo na graduação em medicina.

No que diz respeito à educação médica, a capacidade da IA para modificar e personalizar o processo de aprendizagem é uma das possibilidades futuras mais evidentes nesta área. A inteligência artificial pode contribuir para atender às necessidades específicas dos alunos em meio ao contexto crescente da digitalização, uma vez que possuem preferências e estilos de aprendizagem distintos. A IA pode fornecer conteúdo instrucional adaptativo adequado às lacunas de conhecimento e velocidade de aprendizagem de cada aluno, analisando grandes quantidades de dados e usando algoritmos de aprendizado de máquina[6].

Contudo, a mencionada contribuição da inteligência artificial no ensino de graduação médica ainda não foi totalmente elucidada na literatura, tornando-se crucial para uma análise mais aprofundada de informações relevantes sobre o tema. Por esse motivo, realizamos uma revisão de escopo para mapear a pesquisa científica em dois eixos: o uso de inteligências artificiais como ferramentas de apoio ao ensino-aprendizagem na graduação médica e sua proficiência em diversos componentes curriculares, permitindo identificar lacunas de conhecimento e seus indicadores de evidência, além de sumarizar e disseminar as principais descobertas científicas.

Objetivos

Mapear a literatura científica sobre o uso de inteligências artificiais como ferramentas de suporte ao ensino-aprendizagem na educação médica de graduação, bem como sua proficiência nos componentes curriculares, a fim de identificar lacunas de conhecimento e indícios de evidência.

Métodos

O presente estudo consiste em uma revisão de escopo, cuja elaboração seguiu o *checklist* PRISMA-ScR[7]. A opção da revisão de escopo como método de pesquisa justifica-se por permitir a identificação, análise e investigação de evidências e lacunas científicas, bem como a compreensão das características, conceitos e definições teóricas do campo de conhecimento proposto[8].

Fonte de dados e pesquisa

Para a realização do estudo, foram seguidas 5 etapas: 1 - elaboração da pergunta de pesquisa seguindo o método População (P), Contexto (C), Conceito (C) – PCC; 2 - seleção das bases de dados a serem utilizadas e definição dos termos e estratégias de busca correspondentes; 3 - exportação dos estudos recuperados para o gerenciador Rayyan e estabelecimento dos critérios de inclusão e exclusão; 4 - seleção dos artigos resgatados por 2 revisores blindados/independentes, que seguiram os critérios em prol de uma maior qualidade de evidência; 5 - construção da planilha com a síntese das evidências encontradas, em conformidade com o modelo JBI Manual for EvidenceSynthesis[9].

A pergunta de pesquisa, formulada seguindo o método PCC, foi estabelecida da seguinte maneira: "Em que medida a adoção de inteligências artificiais (IAs) como ferramentas de apoio à educação médica de graduação dinamiza a eficácia pedagógica de docentes e a experiência de aprendizado dos estudantes?" P (população) — docentes e estudantes de cursos de graduação em medicina; C (conceito) — adoção de IAs na dinamização da eficácia pedagógica e experiência de aprendizado; C (contexto) — educação médica de graduação.

Para identificar estudos potencialmente relevantes, foram eleitas as seguintes bases da literatura científica: PubMed, Scientific Electronic Library (SciELO), Web of Science via coleção principal (Clarivate Analytics), Scopus (Elsevier) e Biblioteca Virtual em Saúde (BVS). Em seguida, os termos de busca foram definidos com base no catálogo da National Library of Medicine (NIH), que compila os *Mesh Terms* adequados ao tema. Tais termos, associados aos operadores booleanos, compuseram a seguinte estratégia de busca: ("Inteligência Artificial" OR "AI" OR "Aprendizado de Máquina" OR Chatbot "OU "ChatGPT") AND ("Educação Médica").

Seleção dos estudos e critérios de inclusão e exclusão

As pesquisas recuperadas nas 5 bases de dados foram exportadas para o gerenciador de referências Rayyan, uma ferramenta desenvolvida pelo *Qatar Computing ResearchInstitute* (QCRI) para dinamizar a seleção dos manuscritos. Concomitantemente, os pesquisadores estabeleceram os critérios de inclusão e exclusão que orientaram a etapa de seleção. Dessa forma, foram considerados para inclusão neste estudo manuscritos publicados entre os anos de 2000 e 2023, em qualquer idioma, oriundos de investigações primárias que abordassem a aplicação e/ou análise do uso de inteligências artificiais como suporte ao ensino-aprendizagem na educação médica de graduação. Além disso, foram incorporados estudos que avaliaram a proficiência dessas ferramentas nos componentes curriculares da graduação e no desempenho de atividades docentes. Em contrapartida, foram excluídos artigos que abordavam estudantes de outras áreas, bem como estudos teóricos, revisões, editoriais e artigos de opinião, devido ao seu nível reduzido de evidência científica.

Para reforçar a solidez da pesquisa, dois avaliadores blindados/independentes (D.R.S.G.A; D.G.M.) seguiram um processo de 4 etapas para examinar as publicações recuperadas: inicialmente, foram removidas as duplicações de artigos entre as bases de dados, a fim de evitar erros de análise; em seguida, foram utilizados os filtros do Rayyan para excluir todos os estudos de revisão; logo após, foram analisados os títulos e resumos dos estudos remanescentes, excluindo aqueles que não atendiam aos critérios definidos; e, por fim, foi realizada a leitura na íntegra dos manuscritos cujos títulos e resumos não eram conclusivos em relação aos critérios de elegibilidade, incluindo nesta pesquisa aqueles que cumpriam o objetivo da revisão. As eventuais discordâncias entre os dois avaliadores foram solucionadas por um terceiro pesquisador (R.V.M.), chegando ao conjunto dos artigos que compuseram o presente estudo. Tais etapas de seleção e suas respectivas quantidades de artigos associados foram organizadas graficamente em um fluxograma (Figura 1).

Extração e análise de dados

Após concluir a seleção, os dois avaliadores independentes desenvolveram um formulário de mapeamento de dados para identificar as variáveis a serem extraídas dos estudos revisados. Em um processo interativo de registro e atualização do formulário, as informações relacionadas ao uso de inteligências artificiais como suporte ao ensino-aprendizagem na educação médica de graduação foram organizadas em um quadro Word, com quaisquer divergências sendo resolvidas pelo pesquisador sênior, H.R.S.S. As variáveis registradas seguiram as diretrizes do JBI Manual for Evidence Synthesis[9], abrangendo aspectos como autoria, ano de publicação, país de origem, objetivos, desenho do estudo, tamanho amostral e principais desfechos, os quais compuseram o quadro de evidências do presente estudo (Anexo 1), contemplado nos materiais suplementares. Ademais, elaboramos os gráficos apresentados nesta revisão através do Word.

Resultados

Durante a fase de identificação das publicações, foram recuperados um total de 3.737 manuscritos, oriundos de diversas bases de dados conceituadas. Essas fontes incluíram 1.491 da PubMed, 7 da SciELO, 348 da Web of Science, 1.203 da Scopus e 688 da Biblioteca Virtual em Saúde. Em seguida, utilizando os filtros disponíveis no Rayyan, excluímos 1.174 estudos duplicados, resultando em 2.563 citações. Destes, 2.442 manuscritos foram removidos após análise dos títulos e resumos, restando 121 artigos. Dentre este quantitativo remanescente, 72 foram excluídos por apresentarem desenhos de estudo que não estavam alinhados aos critérios desta revisão, 20 por não abordarem o uso prático de inteligências artificiais na educação médica de graduação e 5 por envolverem estudantes de outros cursos. Ao final, restaram 24 artigos que foram considerados elegíveis para inclusão nesta revisão. O supracitado processo de triagem foi registrado por meio de um fluxograma (Figura 1).

Na elaboração desta revisão de escopo, a Índia destaca-se como origem primária dos estudos selecionados (n=5)[10-14], seguido pelos EUA[15-17] e Canadá[18-20], ambos com n=3 cada. Posteriormente, Alemanha[21,22] e Coreia do Sul[23,24] (n=2). Outros países representados incluem Austrália[25], Arábia Saudita[26], Singapura[27], Dinamarca[28], França[29], Tailândia[30], Taiwan[31], Turquia[32] e Sérvia[33], todos com 1 estudo cada.

Quanto aos tipos de estudos, mais da metade consiste em Estudos Transversais (n=13), o que representa 54,16% do total de manuscritos[10-16,18,21-23,26,32,33]. Dentre estes, 92,30% são estudos transversais quantitativos[10-14,16,18,21-23,26,32,33], e 7,69% são estudos transversais qualitativos[15]. A seguir, são observados Ensaios Controlados Randomizados (n=5), representando 20,83% dos manuscritos[20,24,28,29,31]. Por fim, Estudos de Métodos Mistos (n=3) compreendem 12,50% dos artigos[17,25,30] e Estudos de Coorte (n=2) contabilizam 8,33% das citações[19,27].

Dessa forma, os vinte e quatro estudos investigaram o uso de Inteligências Artificiais (IAs) como ferramentas de suporte ao ensino-aprendizagem na educação médica de graduação. Com o intuito de facilitar a compreensão das diversas aplicações de Inteligências Artificiais nesse contexto, os manuscritos foram classificados e organizados em cinco categorias. A 1ª categoria abrange os estudos em que as Inteligências Artificiais foram utilizadas no treinamento prático dos estudantes, como em ensaios de procedimentos cirúrgicos; a 2ª categoria reúne estudos em que as IAs foram utilizadas na execução de atividades docentes, como na elaboração de questões; a 3ª categoria agrupa os manuscritos que testaram a proficiência das IAs nos componentes curriculares da graduação médica, como na resolução de questões de diferentes disciplinas; a 4ª categoria engloba as pesquisas que avaliaram uso das inteligências artificiais na produção de informações científicas, em comparação com as fontes tradicionais de evidência; e, por fim, a 5ª categoria concentra os estudos que abordaram o impacto dessas ferramentas na aquisição de conhecimento e desempenho acadêmico dos discentes, a partir de uma análise qualitativa e quantitativa. Para esquematizar essa classificação, elaborou-se uma tabela com a relação de artigos correspondentes a cada categoria (Tabela 1).

Ademais, vale ressaltar que, entre as ferramentas de IA mencionadas pelos manuscritos, as que mais se destacaram foram o ChatGPT e suas versões[10-15,18,21-23,26,32,33], sendo objeto de pesquisa em 54,16% dos estudos. Depois disso, a Microsoft Bing[10,12,21] e Google Bardo[10,12,32] foram citados em 12,50% dos artigos cada. Já o instrumento VOA (Assistente Operativo Virtual)[19,20] e outros modelos de *Machine Learning* (ML)[17,25] foram citados, cada um, em 8,33% das referências. Por último, A-LRYNGO[24], AmbuBPS[28], COMETA[30], ChatScript[16], Progresso do bate-papo[29], HipGuide[31] e VIP (*Virtual Integrated Patient*)[27] foram, individualmente, referenciados em 4,16% das fontes. O número de estudos que abordaram cada inteligência artificial pode ser observado na Figura 2, e relatamos a distribuição dessas IAs entre as categorias supracitadas na Tabela 2.

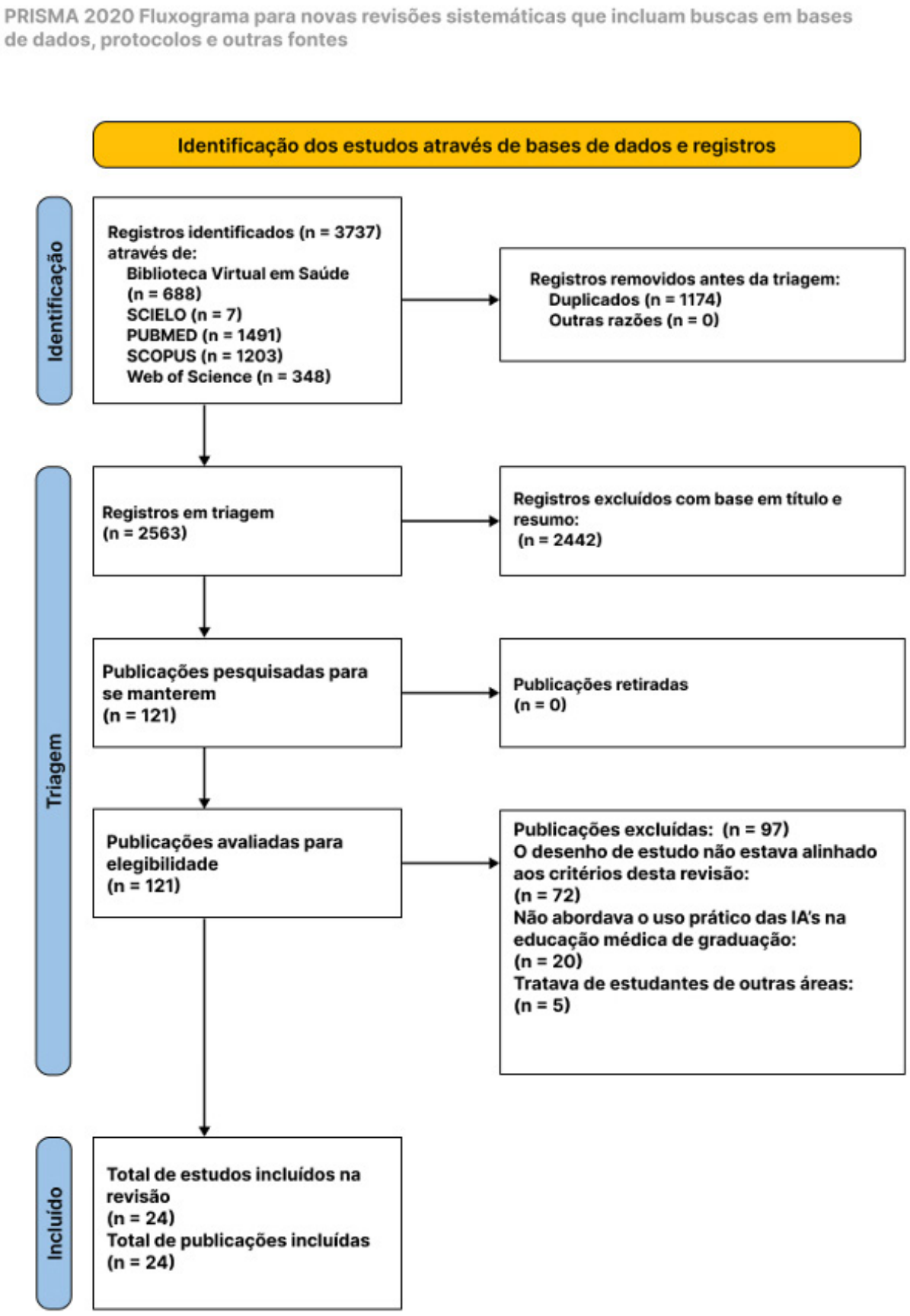

Figura 1. Identificação dos estudos por meio de bases de dados e registros (fluxograma de seleção dos manuscritos).

Fonte: Elaborados pelos autores.

Tabela 1. Distribuição dos artigos revisados por categoria analisada.

1ª Categoria: Uso de IAs no treinamento prático dos discentes	15, 16, 20, 24 e 27.
2ª Categoria: Emprego das Inteligências Artificiais na realização de atividades docentes	8, 12, 13, 21, 26 e 28.
3ª Categoria: Proficiência das Inteligências Artificiais nos componentes curriculares da graduação médica	6, 7, 9, 10, 14, 17, 18, 19, 22, 28 e 29.
4ª Categoria: Uso de inteligências artificiais na produção de informações científicas, em comparação com as fontes tradicionais de evidência	11 e 28.
5ª Categoria: Impacto das IAs na aquisição de conhecimento e desempenho acadêmico dos discentes, a partir de uma análise qualitativa e quantitativa	23, 24 e 25.

Fonte: Elaborados pelos autores.

Tabela 2. Distribuição das IAs por categoria analisada.

1ª Categoria: Uso de IAs no treinamento prático dos discentes	Virtual Operative Assistant (VOA), AmbuBPS, A-LRYNGO, HipGuide.
2ª Categoria: Emprego das Inteligências Artificiais na realização de atividades docentes	COMET, ML, ChatScript, ChatGPT, Microsoft Bing e Google Bard.
3ª Categoria: Proficiência das Inteligências Artificiais nos componentes curriculares da graduação médica	ChatGPT, ChatGPT-4, ChatGPT-3.5-Turbo, Microsoft Bing e Google Bard.
4ª Categoria: Uso de Inteligências Artificiais na produção de informações científicas, em comparação com as fontes tradicionais de evidência	ChatGPT e Google Bard.
5ª Categoria: Impacto das IAs na aquisição de conhecimento e desempenho acadêmico dos discentes, a partir de uma análise qualitativa e quantitativa	Chatprogress, Paciente Virtual Integrado (VIP) e AmbuBPS.

Fonte: Elaborados pelos autores.

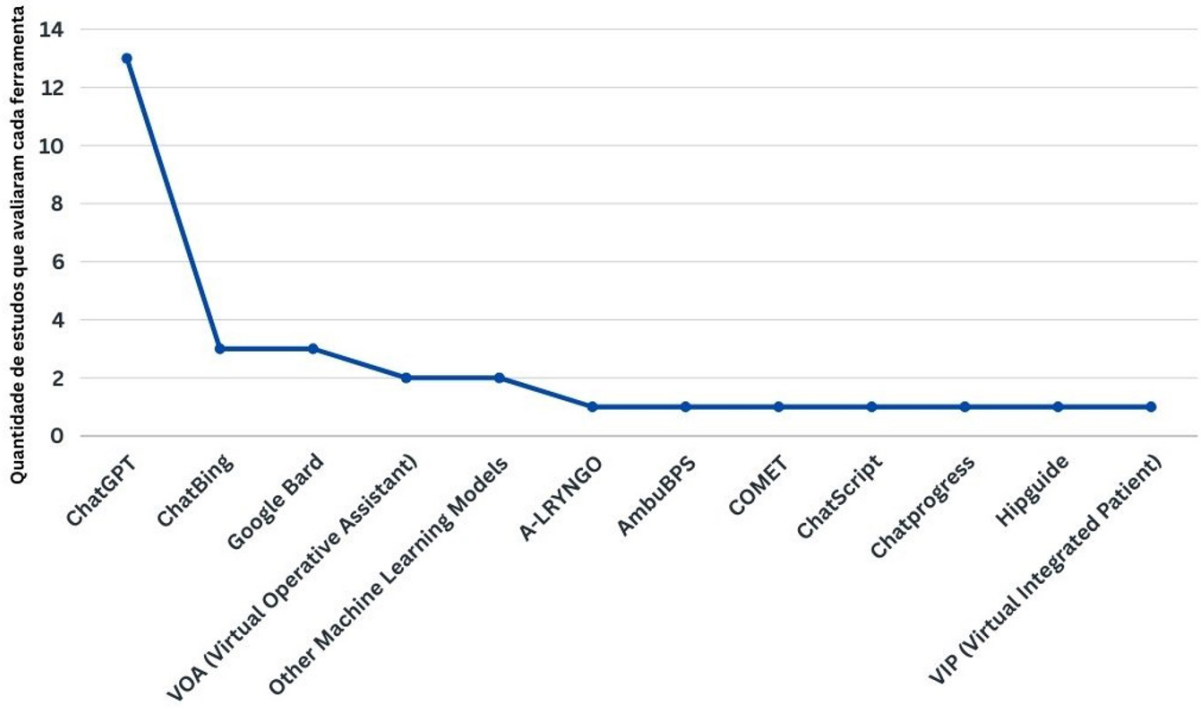

Figura 2. Quantidade de estudos que analisaram, de forma individual ou combinada, cada ferramenta de Inteligência Artificial abordada na presente revisão

Fonte: Elaborados pelos autores.

1ª categoria: utilização da inteligência artificial na formação prática de alunos

Entre os estudos, 20,83% ressaltaram o emprego da IA no treinamento prático dos alunos, considerando parâmetros como maior precisão e qualidade técnica, redução do tempo de execução e segurança na execução dos procedimentos[19,20,24,28,31]. Nesta análise, observou-se que o *Virtual Operative Assistant* (VOA) — um sistema de tutoria inteligente — foi capaz de melhorar a precisão e o desempenho dos alunos do grupo teste durante uma cirurgia de ressecção tumoral, pois sua aplicação foi associada a uma redução considerável na taxa de remoção de tecido cerebral saudável durante a cirurgia, indicando uma melhoria na segurança do procedimento[19]. Outro aspecto relevante é que, após casos de ressecção subpial, o *feedback* da inteligência artificial (VOA) pode ser mais eficaz para melhorar o desempenho em procedimentos cirúrgicos simulados do que a instrução por especialistas ou a ausência de *feedback*[20]. No entanto, é válido salientar que ocorreram algumas alterações não intencionais, como uma diminuição na velocidade e aceleração da mão dominante, o que pode exigir uma avaliação mais aprofundada[19].

De maneira semelhante, outras inteligências artificiais, como AmbuBPS, A-LRYNGOe HipGuide também proporcionaram resultados positivos na melhoria do aprendizado e do treinamento prático de broncoscopias, laringoscopias e análise de fraturas, respectivamente[24,28,31]. O AmbuBPS é um sistema que utiliza inteligência artificial e foi projetado para auxiliar em treinamentos médicos. A tecnologia inclui um escopo de uso único e um conjunto de recursos em tempo real, como rótulos na tela, um diagrama da árvore pulmonar e um escore de progressão estruturada (SP). Esses recursos visam a melhorar a prática de broncoscopia, fornecendo *feedback* em tempo real sobre o posicionamento e o reconhecimento de segmentos bronquiais. No estudo em questão, o AmbuBPS proporcionou resultados superiores em comparação ao grupo controle, especialmente na coordenação da dinâmica, tempo de prova e análise de sequência padrão, permitindo aos discentes sentirem-se mais seguros com o procedimento e mais envolvidos no aprendizado da broncoscopia[28]. No contexto da técnica da laringoscopia, a orientação da glote com base em inteligência artificial utilizando o A-LRYNGO — um videolaringoscópio tipo canal acoplado a um sistema de inteligência artificial — foi promissora, alcançando uma orientação precisa da maioria dos participantes. O estudo destacou que o treinamento prático contribuiu para a melhoria do desempenho dos alunos, evidenciando diferenças notáveis entre os diversos laringoscópios e ressaltando a influência positiva da inteligência artificial na orientação da glote[24]. Em relação ao uso da IA HipGuide — um sistema de IA para detecção de fratura de quadril — na ortopedia, os participantes foram divididos em dois grupos: AIL (grupo teste) e CL (grupo controle). No grupo CL, a pontuação pós-aprendizagem não apresentou diferenças significativas em comparação com a pré-aprendizagem, enquanto no grupo AIL, todos os alunos registraram melhorias significativas com o suporte da IA na interpretação e análise de fraturas em quadril, resultando em pontuações consideravelmente superiores às da pré-aprendizagem[31].

2ª categoria: utilização da inteligência artificial nas atividades docentes

Outro aspecto analisado foi o uso das inteligências artificiais na realização das atividades docentes, como tutoria, elaboração de questões ou mesmo para avaliação das atividades dos estudantes[12,16,17,25,30,32], totalizando 25% dos estudos. Nesse parâmetro, pode ser visto que IAs como COMET e ferramentas de ML conseguiram se equiparar ou mesmo superar tutores humanos e o corpo docente na avaliação dos estudantes em atividades específicas[17,30]. O sistema de tutoria inteligente COMET demonstrou eficácia equiparada a tutores humanos para o acompanhamento e avaliação de estudantes de medicina submetidos ao método de ensino PBL[30]. De maneira similar, a aprendizagem de máquina (ML) revelou-se promissora na análise formativa do raciocínio clínico escrito de estudantes de medicina, sendo capaz de avaliar ensaios de justificação diagnóstica de maneira tão ou mais confiável do que as avaliações do corpo docente[17]. Por outro lado, o ChatScript não apresentou a mesma eficácia nesse tipo de tarefa, tendo um desempenho inferior ao do ser humano na avaliação do desempenho dos estudantes de medicina nos exames clínicos objetivos estruturados (OSCEs), testes feitos com o intuito de avaliar as habilidades comunicativas no contexto de conversas clínicas e entrevistas médicas, conforme pôde ser observado em outro estudo[16]. Ademais, no aspecto qualitativo, verificou-se a divergência entre os estudantes acerca da marcação por computador (*Machine Learning*) de exames escritos. Na referida pesquisa, mais da metade dos alunos expressaram dúvidas sobre a capacidade dos computadores em realizar uma marcação suficientemente precisa, embora a maioria concordasse que a marcação por computador seria mais objetiva do que a marcação feita por humanos[25].

No que diz respeito à capacidade de formular questões, o ChatGPT, o Microsoft Bing e o Google Bard demonstraram potencial na execução da tarefa, mas também apresentaram limitações[12,32]. Um dos estudos revisados examinou a habilidade dessas ferramentas de criar perguntas de múltipla escolha sobre fisiologia e destacou o êxito do ChatGPT e do Google Bard em todas as competências de conteúdo avaliadas, diferindo do Bing, que apresentou falhas em duas competências. Contudo, é importante notar que nenhuma das inteligências artificiais conseguiu gerar um número significativo de questões de múltipla escolha que exigissem um alto nível de compreensão do assunto[12]. Em outro estudo que avaliou diferentes versões do ChatGPT e do Google Bard na criação de perguntas de anatomia, alguns dos exemplares gerados pelo ChatGPT foram considerados incorretos pelos avaliadores anatomistas[32].

3ª categoria: proficiência em inteligência artificial nos componentes curriculares da graduação médica

Adicionalmente, buscou-se compreender a proficiência das IAs nos componentes curriculares da graduação médica,

tanto na resolução de questões quanto em estudos de caso. Sob essa ótica, 41,67% dos estudos podem ser agrupados nesse eixo de análise[10,11,13,14,18,21-23,26,32,33]. Nesse contexto, merece destaque o ChatGPT, que se sobressaiu como a inteligência artificial mais eficaz ao lidar com questões de diversas disciplinas, como anatomia e fisiologia, além de demonstrar sua utilidade na aplicação de estudos de casos de fisiologia e em análises bioestatísticas. Com o objetivo de ilustrar a distribuição dos estudos de proficiência nas diversas áreas da graduação, elaborou-se um gráfico (Figura 3), e suas respectivas referências foram organizadas por meio da Tabela 3.

Em um dos estudos, a análise concentrou-se no desempenho de três *Large Language Models* (LLMs): GPT-4, GPT-3.5-Turbo e Bing. Especificamente, foram examinadas as habilidades desses modelos de resolver questões retiradas dos Exames do Estado Médico Alemão de 2022, considerando cenários com e sem a presença de mídia nas perguntas, e comparando os desempenhos das IAs entre si, e em relação aos alunos, nos exames de primavera e outono de 2022. Comparando as inteligências artificiais entre si no exame de primavera, o GPT-4 destacou-se ao apresentar melhor desempenho na resolução de questões com mídia, seguido pelo Bing e pelo GPT-3.5-Turbo. Quando as questões com mídia foram excluídas, o Bing apresentou melhor desenvoltura na resolução de questões, seguido pelo GPT-4 e pelo GPT-3.5-Turbo. Em comparação com os alunos, o GPT-4 e o Bing apresentaram desenvoltura superior, enquanto o GPT-3.5-Turbo apresentou desempenho comparável ao dos discentes. Comparando as inteligências artificiais entre si no exame de outono, o GPT-4 demonstrou melhor desempenho na resolução de questões com mídia, seguido pelo Bing e pelo GPT-3.5-Turbo. Já quando as questões com mídia foram retiradas, o Bing e o GPT-4 apresentaram desempenhos semelhantes, enquanto o GPT-3.5-Turbo obteve um desempenho inferior. Em comparação com os alunos, o GPT-4 e o Bing mantiveram um desempenho na resolução de questões superior ao dos discentes, porém o GPT-3.5-Turbo obteve resultados inferiores aos dos alunos nesse segundo exame. De modo geral, o GPT-4 apresentou os melhores resultados, seguido pelo Bing com desempenho intermediário e, por último, pelo GPT-3.5-Turbo com a menor desenvoltura[21]. Em outro manuscrito, foi comparada a proficiência das inteligências artificiais na resolução de questões de anatomia em diferentes domínios linguísticos. Com esse intuito, analisaram-se as respostas fornecidas pelas IAs a dois usuários distintos que fizeram as mesmas perguntas. A intenção era verificar se as respostas obtidas em acessos separados eram consistentes ou exibiam variações. Assim, foi revelado que o ChatGPT na linguagem turca alcançou o maior desempenho, seguido pelo Google Bard, também em turco. As demais IAs, incluindo

Tabela 3. Componentes curriculares analisados e seus respectivos estudos.

Multidisciplinar	17, 18, 22, 14.
Fisiologia	6, 9.
Anatomia	28.
Bioestatística	29.
Microbiologia	7.
Bioquímica	10.
Parasitologia	19.

Fonte: Elaborados pelos autores.

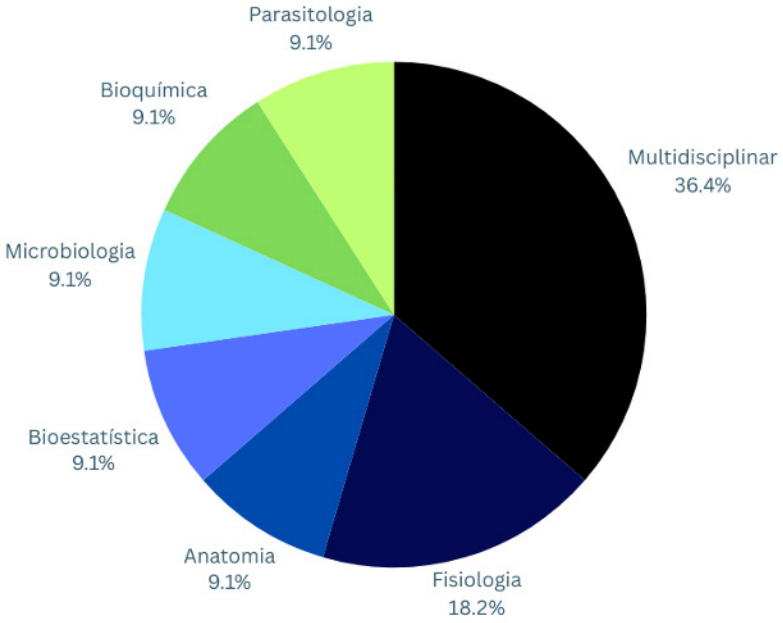

Figura 3. Profissional dos estudos de proficiência por componente curricular analisado.

Fonte: Elaborados pelos autores.

o primeiro usuário do Google Bard em turco, o ChatGPT em inglês e o Google Bard em inglês, apresentaram taxas variadas. Entretanto, nenhuma delas atingiu uma pontuação significativa[32].

Um outro artigo avaliou o desempenho das IAs em estudos de caso de fisiologia, demonstrando que o ChatGPT apresentou consistentemente o melhor desempenho, obtendo pontuações mais elevadas — que foram atribuídas por ambos os avaliadores — em comparação com o Bing e o Google Bard. Os dados detalhados revelam que, em média, o ChatGPT alcançou uma pontuação superior em desempenho entre as três IAs, considerando diferentes sistemas fisiológicos, enquanto o Bing registrou pontuações significativamente mais baixas. O Bard se posicionou como intermediário nessa avaliação[10].

Quanto ao desempenho das Inteligências Artificiais na resolução de questões dos componentes curriculares, é crucial atentar para os resultados divergentes apresentados por um manuscrito, especialmente no que diz respeito à capacidade de solucionar problemas bioestatísticos em diferentes versões do ChatGPT. O GPT-3.5 demonstrou eficácia na resolução de tarefas envolvendo dados categóricos, estudo transversal e teste t para dados pareados, contudo apresentou dificuldades em questões relacionadas ao teste qui-quadrado, análise de variância unidirecional (ANOVA) e tamanho da amostra. Em contraste, o GPT-4 teve um desempenho mais abrangente, conseguindo resolver algumas dessas questões adicionais em tentativas subsequentes[33].

Adicionalmente, é pertinente examinar o desempenho das respostas do ChatGPT em relação aos conteúdos das questões. Em um estudo que envolveu perguntas de múltipla escolha, observou-se que o ChatGPT apresentou pontuações mais baixas em medicina legal e radiologia, contrastando com uma elevada pontuação em dermatologia[22]. Da mesma forma, ao avaliar a eficácia do ChatGPT na resolução de questões fundamentais em fisiologia, destaca-se que a fisiologia cardiovascular alcançou a pontuação mais alta, seguida pela neurofisiologia, enquanto a endócrina registrou a menor pontuação[13]. Em um estudo sobre a habilidade do ChatGPT em responder questões de microbiologia, verificou-se que a categoria de infecções do sistema nervoso central obteve a pontuação mais elevada, enquanto sistemas gastrointestinal e hepatobiliar obtiveram a menor pontuação[11]. Reforçando essas análises, constatou-se que o ChatGPT demonstrou capacidade de responder a questões de ordem superior na área da bioquímica médica, embora seu desempenho tenha sido variável entre as competências da disciplina[14]. Por fim, em outro estudo realizado com o ChatGPT, percebeu-se que ele foi capaz de responder a questões de múltipla escolha nas ciências médicas básicas e clínicas, demonstrando um nível de conhecimento satisfatório. No entanto, o desempenho foi ligeiramente superior nas questões básicas em comparação com as questões clínicas[26].

Por outro lado, em um estudo que comparou as respostas dos alunos às do ChatGPT na resolução de Exercícios de Aplicação de Conceitos (CAEs), as pontuações atribuídas pelos tutores revelaram uma qualidade superior das respostas do ChatGPT em comparação com as dos participantes discentes. Contudo, tanto o desempenho do ChatGPT quanto o dos alunos na pesquisa foram inferiores em relação a um grupo de estudantes previamente submetido ao CAEs[18]. Da mesma forma, em uma análise sobre parasitologia, o desempenho do ChatGPT foi considerado inferior ao dos estudantes de medicina. Algumas limitações identificadas nessa Inteligência Artificial incluíram a incapacidade de interpretar figuras e gráficos, o desconhecimento de dados epidemiológicos específicos da Coreia e a dificuldade em compreender perguntas de múltipla escolha[23].

4ª categoria: utilização da inteligência artificial na produção de informação científica, em comparação com fontes tradicionais de evidência

Além disso, foi avaliada a qualidade das produções científicas, especialmente artigos, geradas pelas inteligências artificiais, em comparação com as fontes de evidência tradicionais, correspondendo a 4,17% dos estudos[15,32]. Um estudo investigou a percepção de estudantes de medicina e de leigos sobre os artigos produzidos pelo ChatGPT, em contraste com as informações de um livro de cirurgia baseado em evidências. Essa pesquisa revelou que os estudantes de medicina consideraram os textos do livro mais abrangentes do que os gerados pelo ChatGPT na abordagem de cada uma das 5 condições cirúrgicas escolhidas. Contudo, os artigos do ChatGPT foram considerados mais claros pelos estudantes de medicina em 4 das patologias cirúrgicas estudadas. Paralelamente, os leigos da população geral notaram uma clareza superior nos artigos do ChatGPT relacionados à apendicite e uma melhor organização nos que abordavam a diverticulite[15]. Em um outro manuscrito deste eixo, foi avaliada a capacidade de redação científica sobre anatomia pelo ChatGPT e pelo Google Bard. Verificou-se que os artigos produzidos por ambas as IA foram menos detalhados e de nível básico em comparação com trabalhos científicos previamente elaborados sobre o tema, além de não apresentarem referências. Contudo, é digno de nota que, entre as duas inteligências artificiais analisadas, o ChatGPT se destacou ao criar um texto mais aderente às normas acadêmicas, apresentando uma estrutura mais refinada e alinhada com a organização padrão requerida para esse tipo de conteúdo[32].

5ª categoria: impacto da inteligência artificial na aquisição de conhecimento e no desempenho acadêmico dos alunos, por meio de uma análise qualitativa e quantitativa

Ademais, sob uma abordagem tanto qualitativa quanto quantitativa, foi mensurado o impacto na aquisição de conhecimento e desempenho acadêmico dos discentes ao utilizarem IAs, correspondendo a 12,50% dos estudos[27-29]. Um desses estudos enfatizou o impacto positivo de jogos interativos que incorporam inteligência artificial, fazendo uso da tecnologia de *Chatbot* (*Chatprogress*), na abordagem de temas de pneumologia. Essa metodologia educacional não apenas superou as expectativas ao aprimorar o desempenho dos

estudantes nos subtestes de pneumologia, mas também trouxe benefícios para a realização do exame global, que abrangia áreas como cardiologia, medicina intensiva e pneumologia. Além disso, dados qualitativos revelaram a alta satisfação dos estudantes com a nova ferramenta de aprendizado, demonstrando interesse contínuo em seu uso[29]. Outro estudo evidenciou que o uso do Paciente Virtual Integrado (VIP) — um simulador virtual de paciente que lança mão dos avanços da inteligência artificial na área de tecnologia de conversação naturalista — ajudou os estudantes de medicina a lembrar os conteúdos metodológicos ensinados nas aulas, aprimorou a compreensão de diferentes casos clínicos, aumentou a confiança e os seus desempenhos na realização de anamneses[27]. Da mesma forma, a aplicação do AmbuBPS na realização de broncoscopias foi bem recebida pelos discentes, que relataram confiança no desempenho durante o procedimento associado ao uso da tecnologia, recomendaram sua utilização e demonstraram interesse contínuo na incorporação dessa ferramenta em seu aprendizado[28].

Discussão

Após análise, as evidências sugerem que as Inteligências Artificiais apresentam um potencial significativo para aprimorar a educação médica de graduação. A versatilidade dessas tecnologias oferece uma ampla gama de aplicações no contexto universitário, podendo servir como uma ferramenta eficaz para apoiar o desenvolvimento dos estudantes de medicina e otimizar o desempenho dos educadores. Em 87,50% dos estudos, sua utilização no ambiente educacional demonstrou ser útil e benéfica de certa maneira, contrastando com apenas 12,50% dos estudos em que as IAs não foram capazes de cumprir as tarefas designadas ou demonstraram desempenho insatisfatório.

1ª categoria: uso de IA na formação prática de alunos

O treinamento prático no curso médico de graduação é fundamental para a formação profissional, pois viabiliza ao aluno o desenvolvimento de habilidades necessárias para a realização de procedimentos (cirúrgicos ou não), a familiaridade com equipamentos médicos e a aplicação do conhecimento teórico desde a formação. Dessa forma, tendo em vista o uso das inteligências artificiais no desenvolvimento de tais aptidões práticas pelos discentes, percebe-se que essas tecnologias são promissoras para aprimorar técnicas práticas e garantir segurança para execução delas. Isso pôde ser constatado com base nos artigos apresentados, a partir do uso de inteligências artificiais como VOA[19,20], AmbuBPS[28], A-LRYNGO[24] e HipGuide[31], as quais viabilizaram a obtenção mais rápida e qualitativa de habilidades pelos discentes. Assim, esses resultados possibilitaram aos estudantes maior segurança em reproduzir tais procedimentos posteriormente, o que minimiza o risco de erros no exercício da profissão. Nesse contexto, esta revisão dialoga com os estudos de Chan et al. (2019)[3] e Narayanan et al. (2023)[34], uma vez que nossos desfechos corroboram as vantagens e as utilidades destacadas por eles em relação aos sistemas instrucionais adaptativos. Essas tecnologias capazes de imitar tutores humanos beneficiam o processo de aprendizado dos discentes, evidenciando a importância da associação das inteligências artificiais no treinamento prático dos estudantes desde a formação.

2ª categoria: utilização da inteligência artificial nas atividades docentes

Além disso, a análise das inteligências artificiais na realização de atividades docentes se mostrou pertinente. Considerando os desafios enfrentados pelos professores universitários, como a carga horária intensa e as múltiplas responsabilidades associadas, incluindo correção de atividades, elaboração de questões e tutorias, o emprego das IAs como uma ferramenta de auxílio para a execução dessas tarefas torna-se uma alternativa promissora. Nesse contexto, tecnologias baseadas em inteligência artificial, como o sistema de tutoria inteligente (COMET) e alguns modelos de aprendizado de máquina (ML), podem ser adotadas como ferramentas para facilitar a avaliação das atividades realizadas pelos alunos. Isso se evidencia a partir dos dados coletados nesta pesquisa, os quais indicam que essas IAs apresentaram desempenho equivalente ou até superior ao humano nessa tarefa. Esta abordagem de inteligência artificial fortalece, portanto, as aplicações destacadas por outros estudos disponíveis na literatura científica[3,34], os quais previamente destacaram a capacidade dos sistemas automatizados de avaliação de redações (AES) de analisar o trabalho dos estudantes, identificar padrões de estudo e conceitos, bem como de fornecer *feedback* individualizado.

Além disso, os *chat boards*, incluindo o ChatGPT, apesar de ainda apresentarem limitações que demandam contínuo aprimoramento dessas IAs, demonstraram potencial como ferramentas de auxílio para os docentes em diversas tarefas, como a elaboração de questões, conforme apontado em estudos que compuseram esta pesquisa.

No entanto, há considerações essenciais a serem ponderadas antes de incorporar oficialmente qualquer IA como ferramenta de ensino-aprendizagem. Uma delas envolve a necessidade de uma triagem cuidadosa para garantir que a inteligência artificial utilizada seja a mais apropriada e confiável. Conforme evidenciado nesta pesquisa, o ChatScript não demonstrou competência suficiente na avaliação do desempenho dos estudantes de medicina nos Exames Clínicos Objetivos Estruturados (OSCEs), destacando que nem todas as IAs são adequadas para desempenhar eficazmente atividades docentes. A outra consideração está relacionada à aceitação do uso dessas ferramentas pelo corpo discente, uma vez que nem todos os alunos confiam plenamente em sua precisão.

3ª categoria: proficiência em inteligência artificial nos componentes curriculares da graduação médica

Outrossim, observou-se que a habilidade das inteligências artificiais em resolução de questões e estudos

de caso, nos diversos componentes curriculares da graduação médica é promissora. Ao tomar como base os resultados dos artigos utilizados nesta pesquisa, percebeu-se que as IAs, como o ChatGPT, por exemplo, podem ser usadas como ferramentas de aprimoramento de aprendizado, pois a solução de questões viabiliza aplicação do estudo teórico, identificação de lacunas no conhecimento, preparação para exames e desenvolvimento de raciocínio clínico. Nesse contexto, destacou-se o ChatGPT tanto para realização de questões quanto para estudos de casos de algumas disciplinas. Além disso, outras versões do ChatGPT e o Bing também foram usadas como instrumento de preparação para exames médicos e mostraram-se ferramentas de estudos eficazes para a resolução de questões e maior preparação pelos alunos.

Contudo, evidenciou-se que questões com informações contidas em mídias não são processadas ou interpretadas pelas inteligências artificiais. Além disso, exercícios envolvendo alguns componentes curriculares, como medicina legal e radiologia, não obtiveram resultados tão satisfatórios e um estudo apontou que as respostas dos alunos foram superiores às do ChatGPT em outra matéria. Com isso, ficou claro que a proficiência das inteligências artificiais nos componentes curriculares ainda apresenta limitações.

4ª categoria: utilização da inteligência artificial na produção de informação científica, em comparação com fontes tradicionais de evidência

Quanto à qualidade das informações científicas geradas por inteligências artificiais, observa-se que, embora os materiais produzidos por elas possuam uma certa qualidade, ainda não conseguem substituir efetivamente as fontes de evidência tradicionais, como livros e artigos convencionais. Apesar da habilidade notável de algumas IAs, como ChatGPT e Google Bard, em gerar artigos de fácil compreensão, seus textos geralmente carecem de profundidade e abrangência em comparação com os presentes em fontes tradicionais. Assim, embora essas IAs possam ser empregadas como ferramentas de busca rápida para auxiliar os alunos de medicina em seus estudos durante a graduação, sua utilização, até o momento deste estudo, não substitui integralmente o uso das fontes de evidências científicas tradicionais como local de consulta mais adequado para o acesso a informações.

5ª categoria: impacto da inteligência artificial na aquisição de conhecimento e no desempenho acadêmico dos alunos, por meio de uma análise qualitativa e quantitativa

Por fim, ficou claro que, além dos aspectos quantitativos relativos aos impactos do uso da inteligência artificial em diversos aspectos da educação médica, os efeitos qualitativos, obtidos a partir de formulários ou questionários, também foram importantes para a compreensão do impacto do uso da IA na aprendizagem dos alunos e desempenho acadêmico. A combinação de dados quantitativos e qualitativos revelou não apenas melhorias tangíveis na aprendizagem, como a simulação de situações realistas e o envolvimento ativo através de jogos (como no *Chatbot*), mas também enfatizou a confiança dos alunos no seu desempenho associado às tecnologias. Esses benefícios associados ao uso da Inteligência Artificial também vão ao encontro das vantagens destacadas no estudo de Narayanan et al. (2023)[34], que enfatiza a utilidade e diversas aplicações dessas tecnologias como ferramentas de apoio à aprendizagem de estudantes de graduação.

Além disso, o interesse contínuo dos estudantes na integração dessas ferramentas, conforme demonstrado nos artigos incluídos nesta investigação, sugeriu que as inteligências artificiais não só melhoraram a aprendizagem, mas também se tornaram elementos recomendados e desejados na educação médica. Assim, esta convergência de resultados reforçou a perspectiva positiva relativamente à integração da IA, indicando um impacto favorável na aprendizagem e no desempenho acadêmico dos estudantes na área da medicina.

Limitações

Esta revisão de escopo apresenta algumas limitações a serem consideradas. Uma limitação recorrente é a focalização em um conjunto específico de grandes modelos de linguagem (LLMs), como ChatGPT, Bard e Bing, o que pode resultar em uma visão limitada da diversidade disponível nesse campo. A avaliação do desempenho desses LLMs frequentemente se baseia em respostas a casos predefinidos, restringindo a generalização para uma variedade mais ampla de cenários clínicos, com preocupações adicionais sobre viés devido aos dados de treinamento. A subjetividade introduzida por avaliações de especialistas, a falta de generalização para outras especialidades médicas e a ausência de considerações sobre potenciais vieses nos LLMs destacam a complexidade dos resultados. Questões críticas incluem a incapacidade de alguns sistemas, como o ChatGPT, interpretarem interações complexas e a limitação na entrada de informações, como gráficos e figuras. O estudo também não abordou completamente potenciais vieses ou erros intrínsecos dos LLMs, apontando para a necessidade de cautela na interpretação dos resultados. Diversos estudos apresentaram desafios metodológicos, como a limitação no número de problemas bioestatísticos investigados e a falta de diretrizes pré-estabelecidas para interações com o ChatGPT, ressaltando a demanda por pesquisas mais abrangentes e refinadas. A revisão destaca que, embora essas pesquisas proporcionem *insights* valiosos, é fundamental reconhecer e abordar essas limitações para uma compreensão mais completa e confiável do papel dos LLMs na educação médica e em outras áreas. Adicionalmente, alguns estudos apresentaram limitações relacionadas à amostragem observada que podem comprometer a generalização e a aplicabilidade dos resultados.

Conclusão

O objetivo desta revisão de escopo foi mapear a literatura científica sobre o uso de inteligências artificiais como ferramentas de suporte ao ensino-aprendizagem na

educação médica de graduação, a fim de identificar lacunas de conhecimento e indícios de evidência. Nesse sentido, foram identificadas 5 áreas principais de aplicação das IAs: 1 - uso de inteligências artificiais no treinamento prático dos discentes; 2 - emprego das inteligências artificiais na realização de atividades docentes; 3 - proficiência das inteligências artificiais nos componentes curriculares da graduação médica; 4 - uso de inteligências artificiais na produção de informações científicas; 5 - impacto das IAs no conhecimento e desempenho acadêmico dos discentes. Os resultados encontrados por este estudo sugerem que a introdução das inteligências artificiais no ambiente acadêmico é benéfica tanto para docentes quanto para os estudantes de graduação, permitindo a dinamização na elaboração de materiais e tutorias, a melhoria em precisão, qualidade técnica e segurança dos procedimentos, além do favorecimento à aquisição de conhecimento e na resolução de questões. Contudo, vale ressaltar algumas limitações identificadas, como a necessidade de aprimoramento na elaboração de questões, desafios na interpretação de gráficos por determinadas inteligências artificiais e a insuficiência na profundidade das informações científicas produzidas. Portanto, esta revisão ressalta a necessidade premente de pesquisas abrangentes e refinadas para alcançarmos uma compreensão mais detalhada do papel das inteligências artificiais na educação médica de graduação. A busca por uma abordagem metodológica mais robusta em estudos posteriores será crucial para validar e generalizar as conclusões, contribuindo, assim, para o avanço efetivo dessa área de estudos.

Contribuição dos Autores

As contribuições dos autores estão estruturadas de acordo com a taxonomia (CRediT) descrita abaixo:

Conceituação & Administração do Projeto: DRSG de Amorim. *Curadoria de dados & Análise formal*: AVNT Xavier; RV de Menezes. *Investigação & Metodologia*: DG Malta; DRSG de Amorim. *Supervisão & Validação*: HR de S e Silva. *Visualização*: DG Malta. *Escrita - primeira redação*: AVNT Xavier; DG Malta; DRSG de Amorim; RV de Menezes. *Escrita - revisão e edição*: AVNT Xavier; DG Malta; DRSG de Amorim; RF de Menezes.

Copyright

Direitos autorais © 2024 Amorim et al. Este é um artigo de Acesso Aberto distribuído sob os termos da Licença Internacional Creative Commons Atribuição 4.0, que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o artigo original seja devidamente citado.

Referências

1. Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. Healthc Manage Forum. 2020;33(1):10-8. DOI: https://doi.org/10.1177/0840470419873123.
2. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem Rev. 2019;119(18):10520-94. DOI: https://doi.org/10.1021/acs.chemrev.8b00728.
3. Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. JMIR Med Educ. 2019;5(1):e13930. DOI: https://doi.org/10.2196/13930.
4. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. Anesthesiology. 2020;132(2):379-94. DOI: https://doi.org/10.1097/ALN.0000000000002960.
5. Ludemir TB. Artificial Intelligence and Machine Learning: Current State and Trends. Estud Av. 2021;35(101):85–94. DOI: https://doi.org/10.1590/s0103-4014.2021.35101.007.
6. Mir MM, Mir GM, Raina NT, Mir SM, Mir SM, Miskeen E, et al. Application of Artificial Intelligence in Medical Education: Current Scenario and Future Perspectives. J Adv Med Educ Prof. 2023;11(3):133-40. DOI: https://doi.org/10.30476/JAMP.2023.98655.1803.
7. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanations. Ann Interm Med. 2018; 169(7):467–73. DOI: https://doi.org/10.7326/m18-085.
8. Mattos SM, Cestari VRF, Moreira TMM. Scoping protocol review: PRISMA-ScR guide refinement. Rev Enferm UFPI. 2023; 12(1). DOI: https://doi.org/10.26694/reufpi.v12i1.3062.
9. Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI. 2020. DOI: https://doi.org/10.46658/JBIMES-20-01.
10. Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. Cureus. 2023;15(8):e42972. DOI: https://doi.org/10.7759/cureus.42972.
11. Das D, Kumar N, Longjam LA, Sinha R, Roy AD, Mondal H, et al. Assessing the Capability of ChatGPT in Answering First- and Second-Order Knowledge Questions on Microbiology as per Competency-Based Medical Education Curriculum. Cureus. 2023;15(3):e36034. DOI: https://doi.org/10.7759/cureus.36034.
12. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023;15(6):e40977. DOI: https://doi.org/10.7759/cureus.40977.
13. Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the Efficacy of ChatGPT in Solving Questions Based on the Core Concepts in Physiology. *Cureus*. 2023;15(8):e43314. DOI: https://doi.org/10.7759/cureus.43314.
14. Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. Cureus. 2023;15(4):e37023. DOI: https://doi.org/10.7759/cureus.37023.

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

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

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

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

19. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in Surgical Curriculum Design and Unintended Outcomes for Technical Competencies in Simulation Training. JAMA Netw Open. 2023;6(9):e2334658. DOI: https://doi.org/10.1001/jamanetworkopen.2023.34658.

20. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. JAMA Netw Open. 2022;5(2):e2149008. DOI: https://doi.org/10.1001/jamanetworkopen.2021.49008.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Este é um artigo de acesso aberto distribuído sob os termos da Creative Commons Attribution License.

Anexo 1. Síntese de evidência.

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Suebnukarn, Haddawy.	2006	Tailândia	120 estudantes de medicina e 10 tutores	Estudo de métodos mistos	Desenvolvimento de técnicas e algoritmos representacionais para gerar dicas de tutoria na resolução de problemas em grupo PBL, bem como a implementação dessas técnicas em um sistema de tutoria inteligente colaborativo, COMET.	Em média, 74,17% dos tutores humanos usaram a mesma dica que o COMET. O tutor humano mais semelhante concordou com o COMET 83% do tempo e o tutor menos semelhante concordou 62% do tempo. Nossos resultados mostram que as dicas do COMET concordam com as dicas da maioria dos tutores humanos com um alto grau de concordância estatística (teste McNemar, p=0,652,κ = 0,773). O foco da atividade do grupo escolhido pelo COMET concorda com o escolhido pela maioria dos tutores humanos com um alto grau de concordância estatística (teste McNemar, p=0,774, κ = 0,823). Os modelos de raciocínio clínico da rede bayesiana podem ser combinados com estratégias genéricas de tutoria para emular com sucesso as dicas do tutor humano no PBL médico em grupo.
Maicher et al.	2019	EUA	Inexistente	Estudo transversal quantitativo	Descrever um sistema virtual padronizado de pacientes (VSP) que permite que os alunos pratiquem suas habilidades de histórico e recebam feedback imediato.	As pontuações gerais calculadas pelo computador não foram diferentes daquelas fornecidas pelos avaliadores humanos, e a precisão geral do sistema de computador foi de 87%, em comparação com 90% para os avaliadores humanos. A confiabilidade entre avaliadores foi alta em 19 de 21 categorias. Desenvolvemos um sistema virtual padronizado de pacientes que pode entender, responder, categorizar e avaliar o desempenho dos alunos na coleta de informações durante um histórico médico típico, permitindo assim que os alunos pratiquem suas habilidades de coleta de história e recebam feedback imediato.
Cheng et al.	2020	Taiwan	30 estudantes de medicina	Ensaio controlado randomizado	Aplicar um sistema de educação de imagem cega prospectiva desenvolvido por eles e confirmar se a IA pode apoiar estudantes de medicina e ajudá-los a aprender com confiança.	O desempenho de pré-aprendizagem foi comparável em ambos os grupos. No grupo CL, a precisão pós-aprendizagem (78,66 ± 14,53) foi maior do que a precisão pré-aprendizagem (75,86 ± 11,36) sem diferença significativa (p=0,64). O grupo AIL mostrou uma melhora notável. A pontuação WithAI (88,87 ± 5,51) foi significativamente maior do que a pontuação de pré-aprendizagem (75,73 ± 10,58, p < 0,01). Além disso, a pontuação pós-aprendizagem (84,93 ± 14,53) foi melhor do que a pontuação pré-aprendizagem (p < 0,01). O aumento na precisão foi significativamente maior no grupo AIL do que no grupo CL. O estudo demonstrou a viabilidade da IA para aumentar a educação médica. A integração da IA na educação médica requer colaboração dinâmica a partir de perspectivas de pesquisa, clínica e educacional.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Cianciolo, LaVoie, Parker	2021	EUA	121 estudantes de medicina e 52 professores clínicos	Estudo de métodos mistos	Avaliar as atitudes dos estudantes de medicina e do corpo docente em relação à IA, em preparação para ensinar fundamentos de IA e aplicações de ciência de dados na prática clínica em um currículo integrado de educação médica.	Apenas 30% dos alunos e 50% dos professores responderam que estavam cientes dos tópicos de IA na medicina. A maioria dos alunos (72%) e do corpo docente (59%) aprendeu sobre IA com a mídia. O corpo docente era mais propenso a relatar que não tinha uma compreensão básica das tecnologias de IA ($\chi 2$, p=.031). Os alunos estavam mais interessados em IA no treinamento de atendimento ao paciente, enquanto o corpo docente estava mais interessado em IA no treinamento de ensino ($\chi 2$, p=0,001). Além disso, estudantes e professores relataram atitudes comparáveis em relação à IA, alfabetização limitada em IA e restrições de tempo no currículo. Há interesse em tópicos amplos e profundos de IA. As descobertas em alunos de medicina e professores estão em paralelo com os resultados da pesquisa de IA de outros grupos profissionais publicados. A pesquisa provou conclusivamente o interesse entre estudantes de medicina e professores na tecnologia de IA em geral e em suas aplicações em saúde e medicina. As pontuações da máquina se correlacionaram mais fortemente com as classificações do corpo docente do que as classificações do corpo docente fizeram umas com as outras (máquina: .28–.53, corpo docente: .13–.33) e foram menos específicas do caso. As pontuações da máquina e as classificações do corpo docente foram igualmente correlacionadas com o conhecimento médico, a cognição clínica e a justificativa diagnóstica prévia. As pontuações das máquinas foram mais fortemente associadas à comunicação clínica do que as classificações do corpo docente (.43 vs .31).
Kong et al.	2021	Cingapura	106 estudantes de medicina	Coorte	Descrever o uso suplementar do VIP na turma do segundo ano, na qual foi realizado um estudo piloto.	Dos 106 participantes, 87% concordaram ou concordaram fortemente que o uso do VIP ajudou a lembrar o conteúdo, enquanto 69% deles sentiram que o VIP aumentou sua confiança e competência na história. O VIP foi bem recebido pelos alunos e mostrou promessa como uma ferramenta para complementar os tutoriais de história, antes do encontro dos alunos com SPs e pacientes reais. Portanto, essa tendência mostrou seu potencial como uma alternativa quando as rotações clínicas foram adiadas ou canceladas.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Choi et al.	2022	Coreia do Sul	30 estudantes de medicina	Ensaio controlado randomizado	Determinar qual dos 4 laringoscópios, incluindo o A-LRYNGO, um videolaringoscópio do tipo canal recém-desenvolvido com um sistema de orientação de glote baseado em inteligência artificial incorporado, é apropriado para o treinamento de intubação traqueal em estudantes de medicina novatos usando equipamentos de proteção individual (EPI).	O tempo para inserir completamente o tubo endotraqueal com o laringoscópio Macintosh não mudou significativamente (p=.177), mas os resultados restantes melhoraram significativamente após o workshop prático (todos p<0,05). Apesar de serem praticantes novatos sem experiência em intubação e usando EPI, os videolaringoscópios do tipo 2 canais foram associados a um bom desempenho relacionado à intubação antes da oficina prática (todos p<0,001). O sistema de orientação de glote baseado em inteligência artificial da A-LRYNGO mostrou 93,1% de precisão, mas 20,7% dos ensaios foram guiados pelas pregas vocais.
Fazllolahi et al.	2022	Canadá	70 estudantes de medicina	Ensaio controlado randomizado	Determinar como a VOA e a instrução remota de especialistas se comparam na aquisição de habilidades, resultados afetivos e cognitivos dos alunos durante o treinamento de simulação cirúrgica.	O ICEMS avaliou 350 ressecções práticas, e o ICEMS e o OSATS avaliaram 70 ressecções realistas. A VOA melhorou significativamente as pontuações de experiência prática em 0,66 (IC 95%, 0,55 a 0,77) pontos em comparação com o grupo de instrutores e em 0,65 (IC 95%, 0,54 a 0,77) pontos em comparação com o grupo controle (p<0,001). As pontuações de especialização realista foram significativamente maiores para o grupo VOA em comparação com o instrutor (diferença de 0,53 [IC 95%, 0,40 a 0,67] pontos; p<0,001) e controle (diferença de valor. 0,49 [IC 95%, 0,34 a 0,61] pontos; p<0,001) grupos. As classificações globais médias do OSATS não foram estatisticamente significativas entre os grupos VOA (4,63 [IC 95%, 4,06 a 5,20] pontos), instrutor (4,40 [IC 95%, 3,88-4.91] pontos) e controle (3,86 [IC 95%, 3,44 a 4,27] pontos). No entanto, nos subpontuações do OSATS, o VOA melhorou significativamente o subpontuponto geral médio do OSATS em comparação com o grupo de controle (diferenda média, 1,04 [IC 95%, 0,13 a 1,96] pontos; p=0,02), enquanto a instrução especializada melhorou significativamente os subpontu os subpontuações do OSATS para manuseio de instrumentos vs controle (diferenda média, 1,18 [IC 95%, 0,22 a 2,14]; p=0,01). Não foram encontradas diferenças significativas na carga cognitiva, ativação positiva e emoções negativas. Portanto, o feedback da VOA demonstrou resultado de desempenho superior e transferência de habilidades, com classificações OSATS equivalentes e respostas cognitivas e emocionais em comparação com a instrução remota de especialistas, indicando vantagens para seu uso em treinamento de simulação.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Agarwal, Sharma, Goswami.	2023	Índia	Inexistente	Estudo transversal quantitativo	Avaliar e comparar a aplicabilidade do ChatGPT, Bard e Bing na geração de MCQs baseados em raciocínio para estudantes de graduação do MBBS (Bacharel em Medicina, Bacharel em Cirurgia) sobre o assunto de fisiologia.	O ChatGPT e o Bard geraram 110 MCQs para as competências escolhidas. No entanto, o Bing forneceu apenas 100 MCQs, pois não conseguiu gerá-los para duas competências. A validade dos MCQs foi classificada como 3 (3-3) para ChatGPT, 3 (1,5-3) para Bard e 3 (1,5-3) para Bing, mostrando uma diferença significativa (p<0,001) entre os modelos. A dificuldade dos MCQs foi classificada como 1 (0-1) para ChatGPT, 1 (1-2) para Bard e 1 (1-2) para Bing, com uma diferença significativa (p=0,006). A capacidade de raciocínio necessária para responder aos MCQs foi classificada como 1 (1-2) para ChatGPT, 1 (1-2) para Bard e 1 (1-2) para Bing, sem diferença significativa (p=0,235). K foi ≥ 0,8 para todos os três parâmetros em todos os três modelos de IA. A IA ainda precisa evoluir para gerar MCQs baseados em raciocínio em fisiologia médica. ChatGPT, Bard e Bing mostraram certas limitações. O Bing gerou MCQs significativamente menos válidos, enquanto o ChatGPT gerou MCQs significativamente menos difíceis.
Al Kahf et al.	2023	França	359 estudantes de medicina	Ensaio controlado randomizado	Avaliar o impacto do Chatprogress, um jogo baseado em chatbot desenvolvido na Universidade Paris Descartes que contém 8 casos de pneumologia, na taxa de sucesso dos alunos em seus exames finais.	As diferenças nas pontuações no subteste de pneumologia ao longo do ano acadêmico foram significativamente maiores entre Jogadores e Usuários vs Controles (pontuação média: 12,7/20 vs 12,0/20, p=0,0104 e pontuação média: 12,7/20 vs 12,0/20, p=0,0365, respectivamente). Essa diferença significativa também estava presente nas pontuações gerais do teste PCC: (pontuação média: 12,5/20 vs 12,1/20, p=0,0285 e 12,6/20 vs 12,1/20, p=0,0355, respectivamente). Embora nenhuma correlação significativa tenha sido encontrada entre as pontuações do subteste de pneumologia e os parâmetros de assiduidade da MS (número de jogos concluídos entre os 8 propostos aos Usuários e número de vezes que um Usuário terminou um jogo), houve uma tendência para uma melhor correlação quando os usuários foram avaliados em um assunto coberto pelo Chatprogress. A MS também foi considerada fã dessa ferramenta de ensino, pedindo mais comentários pedagógicos, mesmo quando acertaram as perguntas. Este ensaio controlado randomizado é o primeiro a demonstrar uma melhora significativa nos resultados dos alunos (tanto no subteste de pneumologia quanto no exame geral de PCC) quando eles tiveram acesso a Chatbots, e ainda mais quando eles realmente o usaram.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Breeding et al.	2023	EUA	56 estudantes de medicina	Estudo transversal qualitativo	Avaliar como estudantes de medicina e leigos avaliam as informações produzidas pelo ChatGPT em comparação com um recurso baseado em evidências sobre o diagnóstico e o gerenciamento de 5 condições cirúrgicas comuns.	Dos 56 participantes da pesquisa, 50,9% (n = 28) eram estudantes de medicina dos EUA e 49,1% (n = 27) eram da população em geral. Estudantes de medicina relataram que os artigos do ChatGPT exibiram significativamente mais clareza (apendicite: 4,39 vs 3,89, *p*=.020; diverticulite: 4,54 vs 3,68, *p*<.001; SBO 4,43 vs 3,79, *p*=.003; sangramento GI: 4,36 vs 3,93, *p*=.020) e melhor organização (diverticulitis: 4,36 vs 3,68, *p*=.021; SBO: 4,39 vs 3,82, *p*=.033) do que a fonte baseada em evidências. No entanto, para todas as 5 condições, os estudantes de medicina descobriram que as passagens baseadas em evidências são mais abrangentes do que os artigos do ChatGPT (colecistite: 4,04 vs 3,36, *p*=0,009; apendicite: 4,07 vs 3,36, *p*=0,015; diverticulite: 4,07 vs 3,36, *p*=0,015; obstrução do intestino delgado: 4,11 vs 3,54, *p*=0,030; sangramento GI superior: 4,11 vs 3,29, *p*=0,003). Os estudantes de medicina perceberam que os artigos do ChatGPT são mais claros e melhor organizados do que fontes baseadas em evidências sobre a patogênese, diagnóstico e gerenciamento de 5 patologias cirúrgicas comuns. No entanto, artigos baseados em evidências foram classificados como significativamente mais abrangentes.
Banerjee et al.	2023	Índia	Inexistente	Estudo transversal quantitativo	Explorar a eficácia do ChatGPT na resolução de várias questões de raciocínio com base nos cinco conceitos principais aplicados a diferentes módulos no assunto da fisiologia.	A pontuação média geral para os módulos (60 perguntas) foi de 3,72 ±0,26, enquanto a pontuação média obtida para os conceitos principais (60 perguntas) foi de 3,68 ± 0,30. Além disso, diferenças estatisticamente significativas (p=0,05 para módulos e *p*=0,024 para conceitos principais) foram observadas entre vários módulos, bem como conceitos principais. As diferenças significativas observadas nas pontuações entre vários módulos e conceitos principais destacam a execução variável da mesma ferramenta de software.
Cold et al.	2023	Dinamarca	20 estudantes de medicina	Ensaio controlado randomizado	Avaliar se o feedback de um sistema automático de identificação de segmento brônquico baseado em IA melhora o desempenho de fim de treinamento dos broncoscopistas novatos	O FG teve um desempenho significativamente melhor em todas as três medidas de resultado (diferença mediana, valor de *P*): completude diagnóstica (3,5 segmentos, *p*<0,001), progresso estruturado (13,5 progressões corretas, *p*<0,001) e tempo de procedimento (–214 segundos, *p*=0,002). O treinamento guiado por esta nova IA faz com que os novatos realizem broncoscopias mais completas, sistemáticas e rápidas.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Das et al.	2023	Índia	Inexistente	Estudo transversal quantitativo	Analisar a capacidade do ChatGPT em responder a perguntas de primeira e segunda ordem sobre o assunto de microbiologia.	A pontuação geral de 96 respostas foi de 4,04 ±0,37 (mediana: 4,17, Q1-Q3: 3,88-4,34), com a pontuação média das respostas às perguntas de conhecimento de primeira ordem sendo 4,07 ± 0,32 (mediana: 4,17, Q1-Q3: 4-4.33) e a das respostas às perguntas de conhecimento de segunda ordem sendo 3,99 ± 0,43 (mediana: 4, Q1-Q3: 3,67-4,33) (Mann-Whitney p=0,4). A pontuação foi significativamente abaixo da pontuação de 5 (teste mediano de uma amostra p<0,0001) e semelhante a 4 (teste mediano de uma amostra p=0,09). No geral, houve uma variação nas pontuações medianas obtidas em oito categorias de tópicos em microbiologia, indicando desempenho inconsistente em diferentes tópicos. Os resultados do estudo indicam que o ChatGPT é capaz de responder a perguntas de conhecimento de primeira e segunda ordem relacionadas ao assunto da microbiologia. O modelo alcançou uma precisão de aproximadamente 80% e não houve diferença entre a capacidade do modelo de responder a perguntas de primeira ordem e perguntas de conhecimento de segunda ordem. Os resultados deste estudo sugerem que o ChatGPT tem o potencial de ser uma ferramenta eficaz para respostas automatizadas a perguntas no campo da microbiologia.
Dhanvijay et al.	2023	Índia	Inexistente	Estudo transversal quantitativo	Avaliar o desempenho de três LLMs, a saber, ChatGPT (3.5; versão de pesquisa gratuita), Google Bard (Experimento) e Microsoft Bing (preciso), na resposta a casos de vinhetas em Fisiologia.	As pontuações gerais para ChatGPT, Bing e Bard no estudo, com um total de 77 casos, foram consideradas 3,19±0,3, 2,15±0,6 e 2,91±0,5, respectivamente, p<0,0001. Portanto, o ChatGPT 3.5 (versão gratuita) obteve a pontuação mais alta, o Bing (Precise) teve a pontuação mais baixa e Bard (Experimento) ficou entre os dois em termos de desempenho. Os valores médios de ICC para ChatGPT, Bing e Bard foram de 0,858 (IC 95%: 0,777 a 0,91, p<0,0001), 0,975 (IC 95%: 0,961 a 0,984, p<0,0001) e 0,964 (IC 95%: 0,944 a 0,977, p<0,0001), respectivamente. O ChatGPT superou Bard e Bing na resposta a vinhetas de casos em fisiologia. Portanto, alunos e professores podem pensar em escolher LLMs para seus fins educacionais de acordo com o aprendizado baseado em casos em fisiologia.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Fazllolahi et al.	2023	Canadá	46 estudantes de medicina e 14 especialistas	Coorte	Avaliar o valor pedagógico das competências técnicas selecionadas pela IA e seus efeitos estendidos no treinamento de simulação cirúrgica.	Um total de 46 estudantes de medicina (idade média [intervalo], 22 [18-27] anos; 27 [59%] mulheres) e 14 cirurgiões (idade mediana [intervalo], 45 [35-59] anos; 14 [100%] homens) foram incluídos neste estudo, e nenhum participante foi perdido para o acompanhamento. O feedback sobre 4 competências técnicas selecionadas pela IA foi associado a mudanças adicionais de desempenho em 32 métricas ao longo de todo o procedimento e 20 métricas durante a remoção do tumor que não foram observadas no grupo de controle. Os participantes expostos ao currículo aprimorado por IA demonstraram melhora significativa nas métricas de segurança, como a redução da taxa de remoção de tecido saudável (diferença −7,05 × 10−5 [IC 95%, −1,09 × 10−4 a −3,14 × 10−5] mm3 por 20 ms; p<0,001) e a manutenção de um controle bimanual focado do campo operacional (diferença mença na divergência máxima do instrumento, −4,99 [IC 95%, −8,48 a −149] mm, p=0,006) em comparação com o grupo de controle. No entanto, efeitos negativos não intencionais também foram observados. Estes incluíram uma velocidade e aceleração significativamente menores na mão dominante (velocidade: diferença média, −0,13 [IC 95%, −0,17 a −0,09] mm por 20 ms; p<0,09]; aceleração: diferença média, −2,25 × 10−2 [IC 95%, −3,20 × 10−2 a −1,31 × 10−2] mm por 20 ms2; p<0,001) e uma redução significativa na taxa de remoção do tumor (diferença média, −4,85 × 10−5 [IC 95%, −7,22 × 10−5 a −2,48 × 10−5] mm3 por 20 ms; p<0,001) em comparação com o controle. Esses resultados não intencionais divergiram as métricas de desempenho de movimento e eficiência dos alunos dos benchmarks de especialização. Neste estudo, um currículo aprimorado por IA para habilidades cirúrgicas bimanuais resultou em mudanças não intencionais que melhoraram o desempenho em segurança, mas afetaram negativamente algumas métricas de eficiência.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Friederichs, Friedrichs, März.	2023	Alemanha	Inexistente	Estudo transversal quantitativo	Medir o nível de conhecimento factual das respostas do ChatGPT, comparando o desempenho do ChatGPT com o dos estudantes de medicina em um teste de progresso.	Das 395 respostas avaliadas, 65,5% das perguntas do teste de progresso respondidas pelo ChatGPT estavam corretas. Em média, o ChatGPT exigiu 22,8 s (SD 17,5) para uma resposta completa, contendo 36,2 (SD 28,1) palavras. Não houve correlação entre o tempo usado e a contagem de palavras com a precisão da resposta ChatGPT (coeficiente de correlação para tempo rho = −0,08, IC 95% [−0,18, 0,02], t(393) = −1,55, *p*=0,121; para contagem de palavras rho = −0,03, IC 95% [−0,13, 0,07], t(393) = −0,54, *p*=0,592). Houve uma correlação significativa entre o índice de dificuldade dos MCQs e a precisão da resposta ChatGPT (coeficiente de correlação para dificuldade: rho = 0,16, IC 95% [0,06, 0,25], t(393) = 3,19, *p*=0,002). O ChatGPT foi capaz de responder corretamente a dois terços de todos os MCQs no nível do exame de licenciamento do estado alemão em Medicina de Teste de Progresso e superou quase todos os estudantes de medicina nos anos 1 a 3.
Ghosh, Bir.	2023	Índia	Inexistente	Estudo transversal quantitativo	Determinar se o ChatGPT pode resolver problemas de ordem superior relacionados à bioquímica médica.	O software de IA respondeu a 200 perguntas que exigiam pensamento de ordem superior com uma pontuação mediana de 4,0 (Q1=3,50, 3D=4,50). Usando uma única amostra de teste de classificação assinada por Wilcoxon, o resultado foi menor que o máximo hipotético de cinco (p=0,001) e comparável a quatro (p=0,16). Não houve diferença nas respostas às perguntas de diferentes módulos CBME em bioquímica médica (Kruskal-Wallis *p*=0,39). A confiabilidade entre classificadores dois professoras. Os resultados desta pesquisa indicam que o ChatGPT tem o potencial de ser uma ferramenta bem-sucedida para responder a perguntas que exigem pensamento de ordem superior em bioquímica médica, com uma pontuação mediana de quatro em cinco.
Hollis-Sando et al.	2023	Austrália	181 alunos de medicina	Estudo de métodos mistos	Avaliar tanto a opinião do aluno em relação à aceitabilidade da pontuação baseada em computador (aprendizado de máquina), quanto um estudo preliminar da precisão da PNL EAD implementada com software de código aberto na pontuação de SAQs de estudantes de medicina.	Embora os alunos tenham expressado preocupação relativamente à precisão da marcação baseada em computador, a maioria dos alunos concordou que a marcação por computador seria mais objectiva do que a marcação humana (67,0%) e relataram que não se oporiam à marcação baseada em computador (55,5%). Em relação à marcação automatizada dos SAQs, para questões de 1 ponto, houve precisões de classificação consistentemente altas (precisão média de 0,98). Para SAQs mais complexos de 2 e 3 pontos, nos quais era necessária classificação multiclasse, a precisão foi menor (média 0,65 e 0,59, respectivamente). Os estudantes de medicina podem apoiar a marcação baseada em computador devido à sua objetividade. A DL tem o potencial de fornecer marcação precisa de questões escritas.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Huh S	2023	Coreia do Sul	77 estudantes de medicina	Estudo transversal quantitativo	Comparar o conhecimento e a capacidade de interpretação do ChatGPT, um modelo de linguagem de inteligência geral artificial, com os dos estudantes de medicina na Coreia, administrando um exame de parasitologia a estudantes de ChatGPT e de medicina.	O ChatGPT respondeu corretamente a 48 de 79 itens (60,8%). Essa pontuação foi menor do que a pontuação média de 77 estudantes de medicina, que foi de 71,8 de 79 (90,8%), com uma pontuação mínima de 65 (89,0%) e uma pontuação máxima de 74 (93,7%). O desempenho do ChatGPT foi menor do que o dos estudantes de medicina. A taxa de resposta correta mostrada pelo ChatGPT não estava relacionada ao nível de conhecimento dos itens. No entanto, havia uma associação entre explicações aceitáveis e respostas corretas. A taxa de resposta correta do ChatGPT de 60,8% não foi necessariamente um indicador de baixo desempenho, pois as perguntas não eram fáceis para os estudantes de medicina responderem corretamente. A pontuação média consideravelmente maior (89,6%) dos estudantes de medicina pode ter sido devido ao seu aprendizado prévio de parasitologia e ao fato de que o exame foi administrado 4 dias após a aula. O conhecimento e a capacidade de interpretação do ChatGPT em responder a este exame de parasitologia ainda não são comparáveis aos dos estudantes de medicina na Coreia.
Ignjatovic', Stevanovic'.	2023	Sérvia	Inexistente	Estudo transversal quantitativo	Avaliar o desempenho do ChatGPT (GPT-3.5 e GPT-4) como uma ferramenta de estudo na resolução de problemas bioestatísticos e identificar quaisquer possíveis desvantagens que possam surgir do uso do ChatGPT na educação médica, particularmente na resolução de problemas bioestatísticos práticos.	O GPT-3.5 resolveu 5 problemas práticos na primeira tentativa, relacionados a dados categóricos, estudo transversal, confiabilidade de medição, propriedades de probabilidade e o teste t. O GPT-3.5 não forneceu respostas corretas em relação à análise de variância, ao teste qui-quadrado e ao tamanho da amostra dentro de 3 tentativas. O GPT-4 também resolveu uma tarefa relacionada ao intervalo de confiança na primeira tentativa e resolveu todas as perguntas dentro de 3 tentativas, com orientação e monitoramento precisos. A avaliação de ambas as versões do desempenho do ChatGPT em 10 problemas bioestatísticos revelou que o desempenho do GPT-3.5 e 4 estava abaixo da média, com taxas de resposta corretas de 5 e 6 em 10 na primeira tentativa. O GPT-4 conseguiu fornecer todas as respostas corretas em 3 tentativas. Essas descobertas indicam que os alunos devem estar cientes de que essa ferramenta, mesmo ao fornecer e calcular diferentes análises estatísticas, pode estar errada, e devem estar cientes das limitações do ChatGPT e ter cuidado ao incorporar esse modelo à educação médica.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Ilgaz, Çelik.	2023	Turquia	Inexistente	Estudo transversal quantitativo	Avaliar como vários recursos de plataformas de IA abertas, como ChatGPT e Google Bard, em sua forma atual, podem contribuir para a educação em anatomia com a perspectiva de questionar, responder e escrever artigos, e visa responder a alguns tópicos controversos.	Os resultados mais bem-sucedidos foram vistos no primeiro usuário do ChatGPT em turco (48,09%), o segundo foi o segundo usuário do Google Bard em turco (45,80%), e os outros foram o primeiro usuário no Google Bard em turco (45,03%), ChatGPT em inglês (44,27%), Google Bard em inglês (41,98%) e o primeiro usuário do ChatGPT em turco (41,22%), respectivamente. Nenhum deles obteve pontuação acima de 50% no total de respostas corretas. Quando os resultados gerais foram comparados em todos os grupos, nenhuma diferença estatisticamente significativa foi encontrada (p>0,05). Houve diferenças no desempenho do LLM, mas nenhuma diferença significativa foi observada entre o ChatGPT e o Google Bard em termos de resposta a perguntas. A questão que gera esses modelos foi satisfeita. Pode ser útil para a preparação do exame em aspectos educacionais. O desempenho dos aplicativos na redação de artigos ainda não estava em um nível suficiente. Embora os LLMs mostrem a promessa como ferramentas valiosas para a educação médica, seu uso requer cautela, dados os algoritmos complexos e suas limitações ao lidar com possíveis imprecisões.
Meo et al.	2023	Arábia Saudita	Inexistente	Estudo transversal quantitativo	Investigar o nível de conhecimento do ChatGPT na educação médica, tanto nas ciências médicas básicas quanto nas clínicas, no desempenho baseado no exame de perguntas de múltipla escolha (MCQs) e seu impacto no sistema de exame médico.	Os resultados revelaram que de 100 MCQs em várias disciplinas de ciências médicas básicas e clínicas, o ChatGPT tentou todos os MCQs e obteve 37/50 (74%) de notas em ciências médicas básicas e 35/50 (70%) em ciências médicas clínicas, com uma pontuação geral de 72/100 (72%) em ciências médicas básicas e clínicas. Conclui-se que o ChatGPT obteve uma pontuação satisfatória em disciplinas de ciências médicas básicas e clínicas e demonstrou um grau de compreensão e explicação. As descobertas deste estudo sugerem que o ChatGPT pode ser capaz de ajudar estudantes e professores de medicina em ambientes de educação médica, uma vez que tem potencial como uma inovação no âmbito das ciências e educação médica.

continua...

... continuação Anexo 1

Autor	Ano	País	Tamanho Amostral	Tipo de Estudo	Objetivo	Desfechos
Morjaria et al.	2023	Canadá	10 alunos de medicina	Estudo transversal quantitativo	Avaliar a extensão da ameaça representada pelo ChatGPT à validade dos problemas de avaliação de resposta curta usados para examinar estudantes de medicina pré-escriação em nosso programa de educação médica de graduação.	As respostas geradas pelo ChatGPT receberam uma pontuação média de 3,29 em 5 (n = 30, 95% IC 2,93-3,65) em comparação com 2,38 para um grupo de alunos que atingiram notas mínimas de aprovação (n = 10, 95% IC 1,94-2,82), representando maior desempenho (p=0,008, $\eta 2$ = 0,169), mas foi superado pelas pontuações médias históricas da classe nos mesmos 30 problemas (média 3,67, p=0,018) ao incluir todas as respostas passadas, independentemente do nível de desempenho do aluno. Não houve uma tendência estatisticamente significativa no desempenho em todos os domínios da Taxonomia de Bloom. Embora o ChatGPT tenha sido capaz de passar em problemas de avaliação de resposta curta que abrangem o currículo pré-escriação, ele superou apenas os alunos com baixo desempenho. Observamos que os tutores em vários casos estavam convencidos de que as respostas produzidas pelo ChatGPT foram produzidas pelos alunos. Os riscos à validade da avaliação incluem incerteza na identificação de alunos com dificuldades e incapacidade de intervir em tempo hábil.
Roos et al.	2023	Alemanha	Inexistente	Estudo transversal quantitativo	Avaliar e comparar o desempenho de 3 LLMs, GPT-4, Bing e GPT-3.5-Turbo, nos Exames Médicos Estaduais Alemães de 2022 e avaliar seu desempenho em relação ao dos estudantes de medicina.	O GPT-4 alcançou o maior desempenho geral, respondendo corretamente a 88,1% das perguntas, seguido de perto pelo Bing (86,0%) e pelo GPT-3.5-Turbo (65,7%). Os alunos tiveram uma taxa média de resposta correta de 74,6%. Tanto o GPT-4 quanto o Bing superaram significativamente os alunos em ambos os exames. Quando as perguntas da mídia foram excluídas, o Bing alcançou o maior desempenho de 90,7%, seguido de perto pelo GPT-4 (90,4%), enquanto o GPT-3.5-Turbo atrasou (68,2%). Houve um declínio significativo no desempenho do GPT-4 e do Bing no exame do outono de 2022, que foi atribuído a uma maior proporção de perguntas relacionadas à mídia e a um potencial aumento na dificuldade de perguntas. Os LLMs, particularmente GPT-4 e Bing, demonstram potencial como ferramentas valiosas na educação médica e para perguntas de exame pré-teste.

Fonte: Elaborados pelos autores.