UNIVERSIDADE PROFESSOR EDSON ANTÔNIO VELANO – UNIFENAS

GUILHERME FREITAS BERNARDO FERREIRA

REVISÃO DE ESCOPO SOBRE O USO DA INTELIGÊNCIA ARTIFICIAL GENERATIVA NO DESENVOLVIMENTO DO RACIOCÍNIO CLÍNICO E A CONFECÇÃO DE GUIA APLICADO AO ENSINO MÉDICO

Belo Horizonte – MG

2025

GUILHERME FREITAS BERNARDO FERREIRA

REVISÃO DE ESCOPO SOBRE O USO DA INTELIGÊNCIA ARTIFICIAL GENERATIVA NO DESENVOLVIMENTO DO RACIOCÍNIO CLÍNICO E A CONFECÇÃO DE GUIA APLICADO AO ENSINO MÉDICO

Dissertação apresentada à Universidade Professor Edson Antônio Velano – UNIFENAS, como parte das exigências do Mestrado Profissional em Ensino em Saúde, para obtenção do título de Mestre Profissional.

Orientadora: Profª. Drª. Maria Aparecida Turci.

Linha de pesquisa: Raciocínio Clínico

Belo Horizonte – MG

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Unifenas BH Itapoã

Ferreira, Guilherme Freitas Bernardo.
Revisão de escopo sobre o uso da inteligência artificial generativa no desenvolvimento do raciocínio clínico e a confecção de guia aplicado ao ensino médico. [Manuscrito] / Guilherme Freitas Bernardo Ferreira. – Belo Horizonte, 2025.
68 f.

Orientadora: Maria Aparecida Turci.

Dissertação (Mestrado) – Universidade Professor Edson Antônio Velano, Programa de Pós-Graduação Stricto Sensu em Ensino em Saúde, 2025.
1. Educação Médica - Raciocínio. 2. Inteligência artificial - Aplicações médicas. I. Ferreira, Guilherme Freitas Bernardo. II. Universidade Professor Edson Antônio Velano. III. Título.

CDU: 61:378

Bibliotecária responsável: Gisele da Silva Rodrigues CRB6 - 2404

Presidente da Fundação Mantenedora - FETA

Larissa Araújo Velano Dozza

Reitora

Maria do Rosário Velano

Vice-Reitora

Viviane Araújo Velano Cassis

Pró-Reitor Acadêmico

Danniel Ferreira Coelho

Pró-Reitora Administrativo-Financeira

Larissa Araújo Velano Dozza

Pró-Reitora de Planejamento e Desenvolvimento

Viviane Araújo Velano Cassis

Diretor de Pesquisa e Pós-graduação

Bruno Cesar Correa Salles

Vice-diretora de Pesquisa e Pós Graduação

Laura Helena Órfão

Coordenador do Curso de Mestrado Profissional em Ensino em Saúde

Aloísio Cardoso Junior

UNIFENAS
Pesquisa e Pós-graduação

Certificado de Aprovação

REVISÃO DE ESCOPO SOBRE O USO DA INTELIGÊNCIA ARTIFICIAL GENERATIVA NO DESENVOLVIMENTO DO RACIOCÍNIO CLÍNICO E A CONFECÇÃO DE GUIA APLICADO AO ENSINO MÉDICO

AUTOR: Guilherme Freitas Bernardo Ferreira

ORIENTADORA: Profa. Dra. Maria Aparecida Turci

Aprovado como parte das exigências para obtenção do Título de Mestre, no Programa de Pós-graduação Profissional de Mestrado em Ensino em Saúde pela Comissão Examinadora.

Profa. Dra. Maria Aparecida Turci

Prof. Dr. Antonio Carlos de Castro Toledo Junior

Profa. Dra. Ligia Maria Cayres Ribeiro

Belo Horizonte 21 de fevereiro de 2025

Prof. Dr. Aloísio Cardos Júnior
Coordenador do Mestrado Profissional
Ensino em Saúde
UNIFENAS

s.castro

Para Nina, Maria Júlia, João Vicente, Maria Cecília e Arthur.

AGRADECIMENTOS

Agradeço à minha mãe, Rosângela, cuja força, dedicação e amor incondicional sempre me motivaram.

À Fernanda, meu amor, companheira de coisas boas e coisas novas.

À minha orientadora Profª. Drª. Maria Aparecida Turci e aos professores Prof. Dr. Alexandre Sampaio Moura e Profª. Drª. Lígia Maria Cayres Ribeiro, pela gentileza de compartilharem suas experiências e se dedicarem a este trabalho.

Aos familiares, representados pelos(as) sobrinhos(as), a quem dedico esta dissertação.

Aos amigos, que trazem a leveza necessária.

À Profª. Drª. Luciana Hoffert Castro Cruz, exemplo de como uma professora muda uma trajetória.

RESUMO

O estudo do raciocínio clínico, compreendido como um processo cognitivo fundamental na prática médica, tem sido objeto de pesquisas desde a década de 1970. A formação dessa relevante competência exige aquisição de conhecimentos, reconhecimento de padrões e desenvolvimento de *scripts* mentais, tornando seu ensino um desafio significativo. Com os avanços da inteligência artificial (IA), particularmente os *chatbots* baseados em modelos de linguagem de larga escala (LLM), como o ChatGPT, novas perspectivas emergem para auxiliar no ensino do raciocínio clínico. Esta dissertação teve como objetivo identificar como os *chatbots* podem ser empregados para a facilitar o ensino do raciocínio clínico. Para tal, foi realizada uma revisão sistemática de escopo, na qual foram selecionadas e analisadas 21 publicações que exploram o uso dessas ferramentas na educação médica. Os resultados sugerem que *chatbot*s podem oferecer suporte ao aprendizado ao gerar listas de diagnósticos diferenciais, fornecer explicações detalhadas sobre casos clínicos e atuar como tutores virtuais em simulações de ensino. A partir desses resultados, foram sugeridas possibilidades de utilização da IA baseadas em princípios educacionais que anteriormente se demonstraram efetivas para o ensino do raciocínio clínico. Também foram identificados desafios, como o risco de viés, a presença de informações imprecisas e a necessidade de supervisão docente no uso dessas ferramentas. Baseado nesses resultados, foi elaborado um guia prático intitulado "6 dicas para usar o ChatGPT para facilitar o ensino do raciocínio clínico", voltado para educadores. Esse material apresenta recomendações embasadas na literatura revisada, incluindo sugestões para otimizar o uso do ChatGPT em atividades pedagógicas. O guia também inclui um glossário com conceitos-chave de IA, tornando-se uma ferramenta acessível para educadores que desejam integrar essa tecnologia ao ensino médico de maneira crítica e eficaz. Dessa forma, este estudo reforça o potencial da IA como aliada no ensino médico, desde que utilizada com planejamento pedagógico e atenção aos desafios inerentes à tecnologia.

Palavras-chave: educação médica; raciocínio clínico; inteligência artificial; tomada de decisões.

ABSTRACT

The study of clinical reasoning, understood as a fundamental cognitive process in medical practice, has been the subject of research since the 1970s. The development of this essential competence requires knowledge acquisition, pattern recognition, and the development of mental scripts, making its teaching a significant challenge. With advances in artificial intelligence (AI), particularly chatbots based on large language models (LLM) such as ChatGPT, new perspectives are emerging to assist in teaching clinical reasoning. This dissertation aimed to identify how chatbots can be used to facilitate the teaching of clinical reasoning. To this end, a systematic scoping review was conducted, in which 21 publications exploring the use of these tools in medical education were selected and analyzed. The results suggest that chatbots can support learning by generating differential diagnosis lists, providing detailed explanations of clinical cases, and acting as virtual tutors in teaching simulations. Based on these findings, AI utilization possibilities were suggested grounded in educational principles that have previously been demonstrated to be effective for teaching clinical reasoning. Challenges were also identified, such as the risk of bias, the presence of inaccurate information, and the need for faculty supervision in using these tools. Building upon these results, a practical guide titled "6 Tips for Using ChatGPT to Facilitate the Teaching of Clinical Reasoning" was developed, aimed at educators. This material presents recommendations based on the reviewed literature, including suggestions for optimizing the use of ChatGPT in pedagogical activities. The guide also includes a glossary of key AI concepts, making it an accessible tool for educators who wish to integrate this technology into medical education in a critical and effective manner. Thus, this study reinforces the potential of AI as an ally in medical education, provided that it is used with pedagogical planning and attention to the inherent challenges of the technology.

Keywords: medical education; clinical reasoning; artificial intelligence; decision-making.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DL	Aprendizagem Profunda
IA	Inteligência Artificial
LLM	Modelos de Linguagem de Larga Escala
ML	Aprendizagem de Máquina
NLP	Processamento de Linguagem Natural
RC	Raciocínio Clínico

SUMÁRIO

1 INTRODUÇÃO

O modo como o pensamento do médico se estrutura para solucionar uma situação clínica é objeto de estudos desde a década de 1970 (Mamede, 2020). Várias pesquisas teorizaram sobre as etapas cognitivas envolvidas, como a coleta e interpretação de dados, a formulação de hipóteses, os processos de julgamento e tomadas de decisão envolvidos na prática profissional. Denomina-se raciocínio clínico (RC) o conjunto de complexas habilidades de aquisição de informações, distinção entre informações relevantes e não relevantes no contexto clínico, geração de explicações para o problema, dedução de quais achados adicionais deveriam ser pesquisados e, por fim, relacionar com as explicações geradas (Mamede, 2020). Compreende-se, portanto, que o raciocínio clínico é uma competência determinante da performance do médico e, dele, depende a acurácia diagnóstica e o sucesso terapêutico. Dessa forma, é fundamental que, durante a formação médica, o estudante seja instigado a desenvolver esta competência.

Contudo, desenvolver o raciocínio clínico dos estudantes é uma tarefa desafiadora, visto que ele se baseia numa extensa base de conhecimentos adquiridos previamente e na construção e reconhecimento de *scripts* mentais — isto é, um construto mental que contém informações epidemiológicas, fisiopatológicas e manifestações clínicas de doenças (Bowen, 2006; Mamede, 2020). Embora essa realidade possa parecer frustrante para os educadores, existem alguns princípios que auxiliam esta tarefa, tais como: assegurar ao estudante a prática com problemas clínicos. de modo a facilitar o desenvolvimento de *scripts* de doenças; estimular a comparação de similaridades e diferenças para demonstrar por que dois problemas que a princípio se assemelham são, na verdade, diferentes; e organizar o ensino de modo a comparar/contrastar exemplos, com a finalidade de desenvolver representações mentais de problemas (Mamede, 2020).

Baseadas nesses princípios, certas abordagens instrucionais se mostraram eficazes em favorecer a construção do raciocínio clínico por estudantes de medicina (Mamede; Schmidt, 2023). Uma delas é a autoexplicação, que requer que o estudante evoque seus conhecimentos sobre o tema explicando-o para si mesmo. Uma outra abordagem é a reflexão deliberada, que consiste em

um modelo sistemático e estruturado de reflexão que orienta médicos a revisarem seu diagnóstico inicial, e forçando-os a considerar diagnósticos alternativos. Evidências demonstram que essa abordagem leva ao aumento consistente de acurácia na solução de casos clínicos complexos (Mamede, 2020; Mamede; Schmidt, 2023; Mamede; Schmidt; Penaforte, 2008). Nota-se que essas abordagens podem ser executadas em diversos cenários, em especial em ambiente real de práticas e através do uso de vinhetas clínicas, por meio da análise de casos escritos (Bowen, 2006).

Esse desafio educacional adquire maior complexidade quando é compreendido dentro do cenário atual, no qual paradigmas educacionais têm sido desafiados pelo desenvolvimento e pela incorporação de novas tecnologias, sendo estas responsáveis pela transformação rápida e constante da vida cotidiana. Paralelamente, é sabido que o volume de conhecimento com o qual o profissional precisa lidar cresce exponencialmente e, especificamente na medicina, estima-se que o volume de informação científica disponível dobra a cada 73 dias (Grunhut; Wyatt; Marques, 2021).

Dentre diversas tecnologias disponíveis, o uso da inteligência artificial (IA) — campo da ciência da computação que busca desenvolver sistemas capazes de realizar tarefas que, tradicionalmente, requerem inteligência humana, como o aprendizado, a tomada de decisão e o reconhecimento de padrões (Civaner *et al.*, 2022; Gordon *et al.*, 2024) — tem sido integrado de forma cada vez mais frequente às atividades do dia a dia para apoiar tomadas de decisões, otimizar tarefas e fornecer *insights* em diversas áreas do conhecimento. A revisão de Grunhut *et al.* (2021) indica que a IA é capaz de melhorar o fluxo de trabalho de atenção médica por meio de triagem automatizada, aumentar a produtividade, reduzir erros humanos, promover melhores padrões de atendimento ao paciente, reduzir custos médicos, realizar cirurgias minimamente invasivas e, até, reduzir taxas de mortalidade. Além disso, vasta literatura já identificou enormes potenciais do uso da IA para aprimorar diagnósticos baseados em imagem, lesões de pele e métodos gráficos (Tolsgaard *et al.*, 2023).

Um tipo específico de IA, denominado ferramenta de modelo de linguagem de larga escala ou *large language model* (LLM), é capaz de compreender e operar a partir da linguagem natural humana. Esse tipo de IA utiliza-se de bilhões de parâmetros textuais para prever a melhor

sequência ou ocorrência em uma frase (Shah; Entwistle; Pfeffer, 2023). Em outras palavras, essa capacidade da IA pode ser compreendida como uma ferramenta "autocompletar" superpoderosa. Os LLMs podem ser utilizados em uma interface amigável para a população em geral, como uma conversa em meio eletrônico — os chamados *chats*. Quando um *chat* é programado para operar de forma automatizada, é denominado *chatbot*. Ressalta-se que um *chatbot* pode ou não utilizar-se de LLM. Contudo, para fins de mais clareza, neste texto, o termo "*chatbot*" será utilizado para se referir aos programas de computador que empregam LLMs, como o ChatGPT (OpenAI, EUA) e o Gemini (Google, EUA).

Cinco dias após o seu lançamento, o ChatGPT, *chatbot* da empresa americana OpenAI, adquiriu rápida popularidade e alcançou um milhão de usuários, seguindo para 100 milhões de usuários em pouco mais de 2 meses (Amaral, 2023; Lee, 2023). Como comparação, mídias sociais profundamente incorporadas no nosso dia a dia, como WhatsApp e Facebook, atingiram 100 milhões de usuários em, aproximada e respectivamente, 3,5 e 4,5 anos.

A versão 4 do ChatGPT (GPT-4) foi submetida a diversos testes de competências médicas, como o USMLE, teste requerido para obtenção da licença médica nos Estados Unidos, e foi aprovada com desempenho notável (Nori *et al.*, 2023). Diante dessas constatações, é importante perceber que os médicos se encontram sob pressão para aprender um conjunto de novas habilidades para lidar com a inteligência artificial (Grunhut; Wyatt; Marques, 2021).

Destaca-se que o ensino médico na graduação representa excelente potencial no treinamento em IA, pois permite uma exposição e integração precoce da IA na educação médica, além de ter a capacidade de alcançar a ampla população de alunos de medicina (Lee *et al.*, 2021). Esses alunos têm a percepção de que o uso da IA pode, dentre outras possibilidades, reduzir erros médicos, facilitar o acesso à informação e promover maior acurácia na tomada de decisões (Civaner *et al.*, 2022).

A integração entre IA e a prática médica, embora pareça um caminho natural e transformador, é um desafio que requer cautela, com especial atenção aos aspectos éticos, aos riscos envolvidos para o paciente e à acentuação de desigualdades relacionadas ao acesso às tecnologias disponíveis (OMS, 2021). Não há motivos para crer que o mesmo não é válido para a educação

médica. Embora tenham sido demonstradas diversas possibilidades de uso de IA na educação médica, existem os riscos de aumentar a dependência dos estudantes em relação à IA, que usam essas ferramentas como fonte primária ou única de informação, e a possibilidade intrínseca de viés e erros por parte da própria IA (Lee, 2023). Há, portanto, a necessidade de colaboração estreita e continuada entre educadores médicos, pesquisadores em educação, cientistas de dados e médicos clínicos para permitir o desenvolvimento de ferramentas de IA que possam apoiar o aprendizado com transparência e atenção a vieses e riscos (Tolsgaard *et al.*, 2023). Nesse sentido, Lee (2023) conclui que

> [...] a aplicação do ChatGPT na educação médica tem um potencial significativo para melhorar as experiências de aprendizagem dos alunos e criar um ambiente educacional mais interativo e envolvente. Com sua capacidade de fornecer informações detalhadas, o potencial para simulações interativas e sua utilidade como recurso educacional, o ChatGPT é promissor para muitos aplicativos educacionais. O uso dessa tecnologia, em conjunto com métodos de ensino tradicionais, pode beneficiar alunos e professores. (Lee, 2023, p. 5)

Essas são reflexões que justificam a realização desta dissertação. Ademais, ao longo dos anos como professor de Neurologia na graduação em Medicina, foi possível perceber como estimular e avaliar o desenvolvimento do raciocínio clínico nos estudantes é um desafio constante — uma habilidade essencial, mas difícil de ensinar de forma estruturada. A inteligência artificial surge, então, como uma ferramenta promissora, capaz de personalizar o aprendizado, oferecer *feedback* imediato e simular cenários clínicos complexos. A vivência como docente demonstra que integrar tecnologia ao ensino não é apenas uma tendência, mas também uma necessidade, de maneira a acompanhar as demandas do ensino em saúde no século XXI.

Nesse contexto, o objetivo desta dissertação é identificar os usos potenciais de *chatbots* baseados em LLM no ensino do raciocínio clínico durante a formação médica, considerando sua aplicabilidade prática, benefícios e limitações. Esta dissertação tem, ainda, um segundo objetivo: utilizar essas evidências no desenvolvimento de orientações que possam ajudar o professor médico a utilizar ferramentas de IA na docência.

O volume desta dissertação está estruturado em capítulos, sendo este primeiro a Introdução. No Capítulo 2, Marco Teórico, são abordados os fundamentos teóricos que conceituam e determinam o desenvolvimento do raciocínio clínico e suas respectivas correlações e aplicações com o ensino em saúde. Ainda nessa seção é apresentado um breve panorama histórico e

conceitual sobre inteligência artificial, como ela se desenvolveu até o estado atual, sempre na medida em que se pode correlacionar com o ensino em saúde. Os pormenores técnicos, computacionais e em engenharia relacionados à IA como campo científico fogem do escopo deste trabalho, sendo apresentados apenas de forma sumária, quando necessário para a compreensão de conceitos sobre o tema. Em seguida, o uso da IA para o ensino do raciocínio clínico é tema da produção bibliográfica, reproduzida no Capítulo 3 (Produção Bibliográfica), no qual é apresentada a revisão de escopo, realizada para compreender o estado atual de conhecimento na área e as principais interlocuções que podem ser estabelecidas entre o uso dos *chatbots*, além das melhores práticas do ensino do raciocínio clínico baseadas em evidências. E, para estimular e facilitar o uso das ferramentas identificadas na revisão de escopo, foi elaborado um guia prático para o uso da IA no ensino do raciocínio clínico para educadores em saúde, apresentado como Produção Técnica no Capítulo 4. O guia prático está acompanhado de um glossário de termos, com o objetivo de facilitar a compreensão e a familiarização de conceitos-chave pelo leitor não especialista. Por fim, no Capítulo 5, são apresentadas as considerações finais deste processo de pesquisa.

2 MARCO TEÓRICO

2.1 Raciocínio clínico: definições e panorama histórico

Pode-se compreender o RC como o conjunto de complexos processos cognitivos que os médicos usam para estabelecer um diagnóstico (Daniel *et al.*, 2019; Mamede, 2020). Esses processos incluem: (i) a maneira como se organizam ideias e pensamentos, (ii) se resgatam, de modo automático ou deliberado, conceitos adquiridos e aprimorados ao longo de sua formação, e (iii) a contemplação desde o reconhecimento de padrões e a evocação de exemplos específicos até a aplicação de conhecimentos básicos de ciências biomédicas (Eva, 2005). Assim, o RC não é apenas uma habilidade abstrata, mas uma integração dinâmica de conhecimento, experiência e habilidades cognitivas adaptadas a cada situação clínica específica (Norman, 2005).

De forma ampla, é possível compreender as capacidades gerais de raciocínio a partir da **Teoria dos Dois Sistemas**, segundo a qual há duas formas de pensar: raciocínio não analítico (Sistema 1), e o raciocínio analítico ou hipotético-dedutivo (Sistema 2) (Kahneman, 2011). No contexto do raciocínio clínico, o Sistema 1 é automático e inconsciente, baseado no reconhecimento de padrões adquiridos por meio de experiências prévias. Nesse modelo, o clínico compara a situação atual com casos similares já enfrentados, permitindo decisões rápidas, mas com maior suscetibilidade a vieses e erros em contextos complexos. Por sua vez, o Sistema 2 envolve processos conscientes e sistemáticos, nos quais o médico analisa cuidadosamente a relação entre sinais, sintomas e diagnósticos. Essa abordagem utiliza regras causais e algoritmos diagnósticos para gerar e testar hipóteses (Eva, 2005).

Os primeiros estudos sobre o raciocínio clínico, realizados no início dos anos 1970, mostraram que os médicos formulam hipóteses diagnósticas logo no início da consulta e, a partir dessas hipóteses, coletam informações para confirmá-las ou descartá-las. Esse processo é derivado do raciocínio hipotético-dedutivo (Sistema 2), pois começa com algumas hipóteses preliminares e utiliza recursos como perguntas durante a anamnese e dados do exame físico para testá-las. Todavia, logo essa teoria foi confrontada com a observação de que, de estudantes novatos a médicos experientes, todos demonstraram o mesmo processo cognitivo. Assim, o que definia a

acurácia diagnóstica era o fato de que as hipóteses elaboradas inicialmente pelos *experts* eram melhores. Além disso, observou-se que o sucesso em um problema clínico não predizia o sucesso em outros, mesmo dentro da mesma especialidade. A acurácia diagnóstica dos médicos experientes estava, portanto, relacionada ao conhecimento específico do conteúdo, e não ao processo cognitivo (Custers; Regehr; Norman, 1996; Mamede, 2020; Norman, 2005).

A partir da compreensão de que a *expertise* dependia mais do conhecimento específico do que de habilidades gerais de raciocínio, os esforços nas pesquisas voltaram-se para o conteúdo, isto é, como o conhecimento se estrutura, se organiza e é ativado na memória do *expert* (Mamede, 2020). Evidências em pesquisas de outras áreas do conhecimento, como o xadrez, demonstraram que a capacidade de um *expert* relembrar a posição de peças no meio-jogo era um forte indicador de seu sucesso (Charness, 1992; Simon; Chase, 1988). Baseados nessas evidências, pesquisadores tentaram aplicar o mesmo conceito na medicina. Contudo, as pesquisas demonstraram que, diferentemente do xadrez, no qual lembrar-se da posição de cada peça é crucial, na medicina, acumular e recordar grandes quantidades de dados dos pacientes não determina a acurácia diagnóstica. Soma-se a isso evidências de que a quantidade total de conhecimento de um médico, da mesma forma, não é capaz de explicar isoladamente sua capacidade diagnóstica (Norman, 2005).

Apesar da complexidade da tarefa empreendida, de desvendar o processo cognitivo gerador do raciocínio clínico, fica claro que *experts* apresentam, à disposição, conhecimento em maior quantidade, de tipos diferentes, melhor organizado e mais acessível em comparação aos novatos (Norman, 2005). Segundo Mamede (2020),

> [...] o que acontece no processo diagnóstico é que alguns achados do problema ativam na memória do médico o conhecimento da doença a qual eles estão associados e uma hipótese é gerada, desencadeando uma busca, guiada pela representação mental da doença, de informações para checar se outros achados associados àquela doença estão presentes. (Mamede, 2020, p. 3)

A ideia inicial de que o raciocínio clínico se baseia num processo cognitivo hipotético-dedutivo permanece. Contudo, ele não é um processo independente de conteúdo, mas, sim, uma estratégia para acessar e utilizar conhecimentos organizados em representações de doenças (*scripts* de doenças) armazenados na memória (Mamede, 2020).

A partir do trabalho seminal de Schmidt *et al.* (1992), foi possível compreender que a *expertise* é construída de tal maneira que, à medida que se aumenta a experiência, os médicos transitam de forma contínua entre três tipos de "representações mentais": de mecanismos básicos das doenças, para *scripts* de doenças e, assim, para exemplos derivados da experiência. Com isso, foi possível esclarecer o que é observado na prática durante a trajetória de iniciante a *expert*: no início da formação, o estudante forma, na memória, redes ligando achados clínicos aos processos fisiopatológicos relacionados. Nesse estágio ele ainda não é capaz de reconhecer padrões e explica achados clínicos isolados com base na fisiopatologia. Mais tarde, esses conhecimentos são mobilizados para compreender problemas clínicos, sendo gradualmente "encapsulados". Essas "cápsulas" contém em si diversos conceitos e suas relações — por exemplo, a fisiopatologia, fatores predisponentes/epidemiológicos e consequências clínicas. Essas "cápsulas" são os chamados *scripts* de doenças. A partir desse ponto, a experiência e a prática com pacientes induzem o desenvolvimento e o fortalecimento desses *scripts*, além da incorporação de apresentações atípicas (Mamede, 2020; Norman, 2005; Oliveira *et al.*, 2022).

Os *scripts* de doenças têm, portanto, um papel central no processo diagnóstico. Durante o atendimento, as características do paciente ativam na memória do médico um ou mais *scripts*, e é essa partícula de informação que guia a busca por mais informações para reforçar ou refutar as hipóteses diagnósticas geradas inicialmente (Mamede, 2020; Schmidt; Rikers, 2007). Logo, torna-se claro que o raciocínio clínico é aprimorado na transição de iniciante a *expert*, oscilando por essas etapas fundamentais, isto é, inclui o "encapsulamento" do conhecimento em *scripts* de doenças.

2.1.1 Raciocínio clínico: é possível ensiná-lo?

É consenso que o raciocínio clínico é um componente fundamental dentre várias competências médicas, e educadores concordam que deve ser ensinado para estudantes de medicina. Contudo, diversas pesquisas têm demonstrado reiteradamente que (i) não há uma habilidade de raciocínio genérica que seja carregada de um problema a outro (Mamede, 2020); (ii) a capacidade de raciocínio não é um traço que possa ser simplesmente atribuído a um indivíduo (Eva, 2005). A

evidência atual sugere que o raciocínio clínico de *experts* é a consequência de uma base de conhecimentos extensa e multimodal (Norman, 2005).

Uma vez que o raciocínio clínico evolui por meio da encapsulação do conhecimento biomédico e clínico em conceitos diagnósticos mais simplificados e, posteriormente, na formação de *scripts* de doenças, o foco do ensino do raciocínio clínico deve ser em construir essa base de conhecimentos, e não nos processos cognitivos do raciocínio em si (Schmidt; Mamede, 2015). Partindo desse pressuposto, surgem na literatura princípios que devem ser perseguidos nas estratégias de ensino durante a formação médica.

Primeiro, forte corpo de evidências sugere que assegurar ao estudante a prática a partir de problemas clínicos, que sejam suficientemente complexos na medida em que os estudantes consigam analisá-los, mas que representam um desafio para sua solução, é essencial para o desenvolvimento do raciocínio clínico (Bowen, 2006; Eva, 2005; Mamede, 2020; Schmidt; Mamede, 2015; Schmidt; Rikers, 2007). Nesse sentido, segundo Eva (2005), o benefício da prática com problemas é tanto maior quanto mais realista for a apresentação do caso. Isto é, o exercício diagnóstico deve se assemelhar aos cenários reais, nos quais o diagnóstico é parte do processo de resolução do caso, e não apenas um tema a ser aprendido. É, também, fundamental promover oportunidades de aprendizagem com vários problemas diferentes em contextos mais diversos possíveis, uma vez que os cenários de prática reais (em enfermarias e ambulatórios, por exemplo) podem confrontar o estudante com um número limitado de casos durante a formação (Cooper *et al*., 2021; Schmidt; Mamede, 2015).

Segundo, a ciência biomédica básica deve ser ensinada de forma contextualizada e de maneira que permita o encapsulamento de conceitos. Para isso, a integração entre esses conceitos e o conhecimento clínico deve ser estabelecida o quanto antes (Schmidt; Rikers, 2007). Sugere-se, dessa forma, que essa abordagem pode ser mais eficaz que o modelo Flexneriano tradicional do ensino médico, no qual o ensino se dá em duas etapas: primeiro, o estudante aprende os conceitos biomédicos e, depois, lida com problemas clínicos (Eva, 2005).

Terceiro, a prática com problemas clínicos deve ser executada considerando o nível de *expertise* dos estudantes. Em outras palavras, vale citar Schmidt e Mamede (2015, p. 962): “estudantes

de níveis de treinamento diferentes requerem diferentes níveis de assistência". Embora possa parecer óbvio, ignorar esse princípio prejudica a aprendizagem. Por exemplo, nos estágios iniciais de aprendizagem, apresentar o problema inteiro (*whole-case*) é mais eficaz do que a apresentação mais próxima da realidade. A exposição de todos os dados relevantes em conjunto para a análise do caso, diferentemente do cenário real — no qual o médico precisa antever quais informações devem ser adquiridas —, reduz a carga cognitiva de aprendizagem (Cooper *et al.*, 2021; Schmidt; Mamede, 2015).

Por último, tempo suficiente deve ser designado para a reflexão em todas as estratégias de aprendizagem (Schmidt; Rikers, 2007). Trata-se, aqui, de um princípio salutar da literatura sobre ensino do raciocínio clínico que requer maior foco e atenção.

Portanto, o desafio de ensinar raciocínio clínico a estudantes de medicina requer estratégias pedagógicas que integrem conhecimentos biomédicos e clínicos de forma contextualizada, promovam práticas diagnósticas em cenários realistas e desafiadores, e respeitem os diferentes níveis de *expertise* dos aprendizes (Cooper *et al*., 2021). A ênfase deve estar na construção de uma base sólida de conhecimentos e na prática reflexiva, permitindo que o estudante desenvolva gradualmente a habilidade de encapsular informações e criar *scripts* de doenças (Mamede, 2020; Schmidt; Rikers, 2007). Dessa maneira, o ensino do raciocínio clínico pode ser eficazmente estruturado, alinhando-se às evidências atuais sobre como essa competência fundamental se desenvolve ao longo da formação médica.

2.1.2 Ensino do raciocínio clínico: por que é tão importante "refletir"?

A reflexão promove a organização, o refinamento e a aplicação dos *scripts* de doenças, melhorando a capacidade dos estudantes de diagnosticar com precisão e eficácia em situações clínicas diversas (Cooper *et al*., 2021; Prakash; Sladek; Schuwirth, 2019).

Novamente, é importante ressaltar que as iniciativas educacionais que pretenderam apenas ensinar a teoria dos processos cognitivos ou treinar uma estratégia de raciocínio específica não demonstraram benefício (Schmidt; Mamede, 2015; Sherbino *et al*., 2014). De acordo com

Schmidt e Mamede (2015, p. 968), uma boa explicação para isso é que “estratégias de raciocínio gerais não existem separadamente do conhecimento sobre uma doença em particular. Logo, é infrutífero ensiná-los de maneira isolada”. Em outras palavras, se é uma extensa base de conhecimentos que determina o desempenho diagnóstico, desenvolvê-la deve ser o foco primordial das estratégias educacionais (Mamede, 2020; Schmidt; Mamede, 2015).

Essa premissa é reforçada por uma revisão sistemática e metanálise que demonstraram que intervenções educacionais que promovem a reflexão — guiada ou não — foram as mais eficazes em melhorar a tomada de decisão diagnóstica dentre outras ferramentas, a saber: *workshops* e *checklists* para reduzir vieses cognitivos, instruções para induzir uma abordagem diagnóstica analítica, uso de *feedback* e técnicas mistas (Prakash; Sladek; Schuwirth, 2019).

Para auxiliar o estudante a desenvolver o seu raciocínio clínico é, portanto, fundamental fazê-lo não só puramente refletir, mas refletir com o objetivo de facilitar o processo de formação de *scripts* de doenças na memória (Bowen, 2006; Mamede, 2020). Partindo dos princípios demonstrados no tópico anterior, o sucesso educacional da prática com problemas depende daquilo que os estudantes são instruídos a fazer com os casos (Mamede, 2020).

O consenso elaborado por Cooper *et al* (2021) identifica diversas estratégias de ensino que, em última análise, ajudam a produzir *scripts* de doenças e são eficazes no desenvolvimento do raciocínio clínico em estudantes de medicina. Dentre as principais, destacam-se as estratégias que constroem entendimento, como a autoexplicação, que incentiva os alunos a explicarem em voz alta seus pensamentos e raciocínios. Essa prática auxilia na consolidação do conhecimento, conectando novas informações ao conhecimento prévio (Chamberland *et al*., 2011; Mamede, 2020). Por exemplo, o aluno pode ser instigado a explicar os mecanismos biomédicos e fisiopatológicos subjacentes aos sinais e sintomas, de modo a consolidar seu *script* mental para o problema em análise.

As estratégias de reflexão estruturada também se mostraram relevantes e demonstraram o aumento da performance diagnóstica em estudantes de variados níveis. Pode-se compreender a reflexão estruturada como uma estratégia de ensino que exige que os alunos analisem cuidadosamente um caso clínico, listando os achados que apoiam ou contradizem um

diagnóstico específico, e, em seguida, reflitam sobre diagnósticos alternativos (Mamede *et al.*, 2012). Esse tipo deliberado de reflexão incentiva os alunos a refletirem sobre suas hipóteses diagnósticas e a considerarem alternativas, fazendo perguntas como “Qual é a evidência para isso?” ou “O que mais poderia ser?”. Dessa forma, é possível enriquecer os *scripts* de doenças, o que facilita o processo de diagnóstico quando os alunos encontram novos casos dessas mesmas doenças no futuro.

A experiência com casos clínicos acompanhada de *feedback* corretivo é outra estratégia com eficácia. Isso porque a prática diversificada com diferentes casos e em vários contextos é fundamental para o aprendizado, mas precisa ser acompanhada de *feedback* para ser verdadeiramente eficaz. Assim, estratégias que envolvem o esforço consciente de recuperação ativa de informações melhoram a retenção a longo prazo e, portanto, o desempenho diagnóstico (Cooper *et al*., 2021).

É importante lembrar que todas essas estratégias de ensino devem corresponder ao estágio de aprendizado dos estudantes. Para iniciantes, tarefas de baixa complexidade com alto suporte instrucional são mais apropriadas, enquanto, para alunos mais avançados, tarefas de alta complexidade com menor suporte são mais eficazes (Cooper *et al.*, 2021).

2.2 Inteligência artificial

2.2.1 Definições

A inteligência artificial (IA), como definida anteriormente, é um campo da ciência da computação. Trata-se de um termo abrangente que contempla ampla gama de tecnologias computacionais que incluem, por exemplo: sistemas especializados, visão computacional, robótica e aprendizado de máquina (Park *et al*., 2021). A seguir, alguns conceitos relacionados à IA são esclarecidos com profundidade suficiente para contextualizar como essa ferramenta pode transformar áreas como Medicina e Ensino em Saúde:

a. Redes Neurais: modelo computacional inspirado na forma como as redes neurais biológicas presentes no cérebro humano processam informações. Consiste em grupos interconectados de nós (ou "neurônios") que trabalham juntos para resolver problemas específicos, frequentemente envolvendo reconhecimento de padrões (Lecun; Bengio; Hinton, 2015).

b. Aprendizado de Máquina (*Machine Learning* – ML): conceito-chave para compreender o funcionamento por trás da maioria das funções da IA. Diz respeito à ciência computacional que aborda a questão de como construir programas que se aprimoram automaticamente por meio da experiência (Jordan; Mitchell, 2015).

c. Aprendizado Profundo (*Deep Learning* – DL): subcampo do aprendizado de máquina que utiliza redes neurais artificiais com várias camadas para processar informações e modelar relações complexas entre os dados (Heilbroner; Miotto, 2023).

d. Processamento de Linguagem Natural (*Natural Language Processing* – NLP): qualquer tipo de abordagem computacional ou matemática que lida com a linguagem humana natural (escrita ou falada) (Gordon *et al*., 2024).

e. Visão computacional: tecnologia que utiliza ML e DL para permitir às máquinas reconhecer, interpretar e analisar imagens, vídeos e outras informações visuais.

f. *Chatbot*: programa de computador que simula conversas humanas, permitindo que os usuários interajam com dispositivos digitais, podendo ser controlado por IA, NLP, ML ou por regras programadas.

Os conceitos atuais em IA têm origem a partir do desenvolvimento de ferramentas computacionais que datam da metade do século XX, desde o questionamento — "Máquinas podem pensar?" — proposto por Turing em seu célebre artigo, no qual argumenta que a distinção entre inteligência humana e inteligência artificial torna-se menos relevante caso uma máquina consiga imitar o comportamento humano de forma tão convincente que um

interrogador não consiga diferenciá-la de um interlocutor humano. Para avaliar essa capacidade, ele propõe o **Jogo da Imitação** (Turing, 1950).

Em 1956, pesquisadores se reuniram na Conferência de Dartmouth, período reconhecido como o marco inicial do campo científico da inteligência artificial, com o objetivo de criar uma máquina que pudesse reproduzir todos os aspectos da inteligência humana (McCarthy *et al.*, 1956). Foram inúmeros os avanços tecnológicos e computacionais que permitiram a evolução da IA até os dias atuais, de modo que seu detalhamento foge do escopo desta dissertação. De maneira geral, pode-se dizer que os sistemas baseados em IA evoluíram de sistemas especializados — isto é, desenhados de maneira algorítmica para finalidades específicas — para abordagens mais avançadas, que incluem aprendizado de máquina, redes neurais artificiais e aprendizado profundo. Essas tecnologias possibilitaram o desenvolvimento de modelos mais complexos e adaptativos, permitindo avanços em diversas áreas, como o NLP. Nesse campo, por exemplo, o aprendizado de máquina e as redes neurais são aplicados para interpretar, processar e gerar linguagem humana de forma cada vez mais precisa e sofisticada, abrindo novas possibilidades para interações mais naturais e compreensivas entre humanos e máquinas.

2.2.2 Inteligência artificial na Medicina

O interesse e as primeiras aplicações da IA na medicina datam do início da década de 1970, com o surgimento de grupos de pesquisa que buscavam utilizar tecnologias computacionais em áreas como Química, Biologia e Ciências Biomédicas (Patel *et al.*, 2009). Essas iniciativas incluem o MYCIN (Shortliffe, 1976), um dos primeiros sistemas especializados voltados para o diagnóstico de infecções bacterianas e a recomendação de antibióticos. Esses sistemas utilizavam regras algorítmicas baseadas no conhecimento humano, ajudando a estabelecer a IA como uma ferramenta de apoio ao diagnóstico. Conforme Patel *et al.*:

> A comunidade de pesquisa geral em IA estava fascinada com as aplicações que estavam sendo desenvolvidas no mundo médico, observando que novos métodos significativos de IA estavam surgindo à medida que os pesquisadores enfrentavam problemas biomédicos desafiadores. De fato, em 1978, o principal periódico da área (Artificial Intelligence, Elsevier, Amsterdã) dedicou uma edição especial exclusivamente a artigos

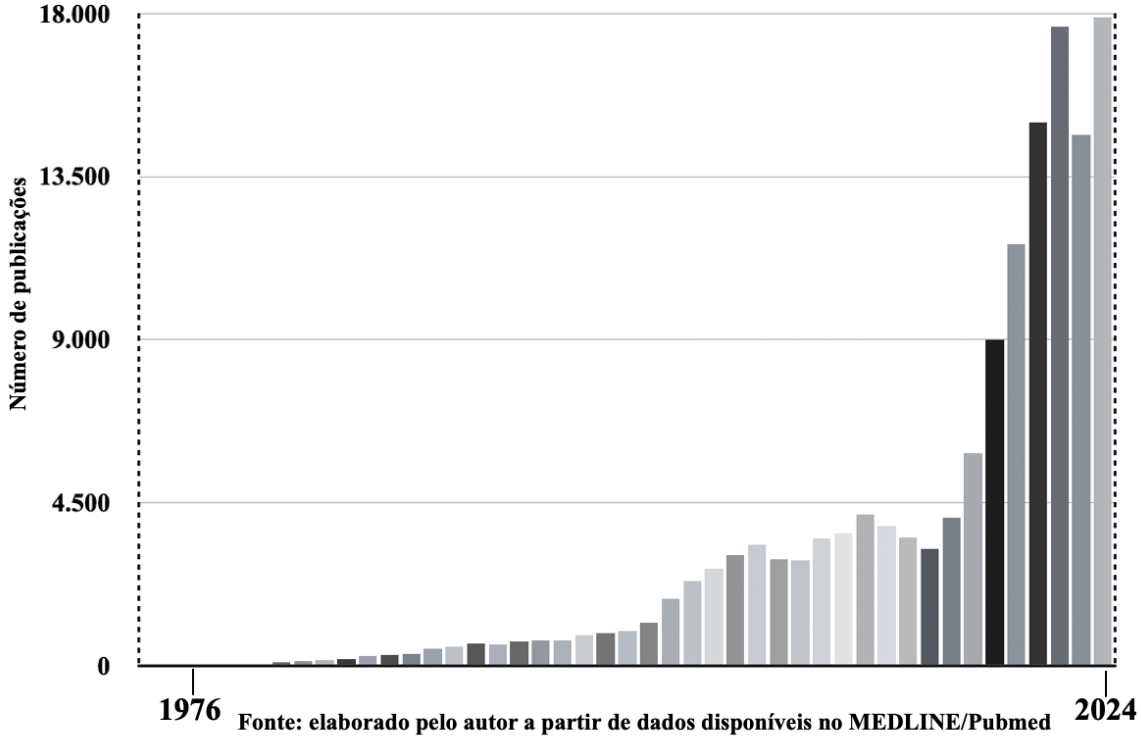
> de pesquisa em IA aplicada à medicina. Ao longo da década seguinte, a comunidade continuou a crescer, e com a formação da Associação Americana para a inteligência artificial em 1980, um subgrupo especial sobre aplicações médicas (AAAI-M) foi criado. (Patel *et al.*, 2009, p. 2)

As aplicações atuais da IA na medicina são diversas e abrangem: (i) facilidades administrativas como o agendamento *online* de consultas, digitalização de registros médicos, telemedicina e atendimentos remotos (Amisha *et al.*, 2019); (ii) ferramentas complexas de análise diagnóstica em áreas como genética médica, radiologia, neurofisiologia e patologia (Jiang *et al.*, 2017); (iii) inovações tecnológicas em dispositivos médicos como Interface Cérebro-Máquina (Lebedev; Nicolelis, 2006).

É reconhecido atualmente que o uso da IA pode auxiliar médicos a tomarem decisões clínicas melhores e, nesse sentido, pode operar de duas formas: a primeira usa ferramentas de ML para analisar grandes volumes de dados estruturados gerados por métodos de diagnóstico por imagem, genéticos e neurofisiológicos; a segunda inclui o uso de NLP para extrair informações não estruturadas de registros clínicos ou da literatura médica e, assim, enriquecer e embasar decisões (Heilbroner; Miotto, 2023; Jiang *et al.*, 2017; Sheikhalishahi *et al.*, 2019).

Inicialmente, o uso da IA na medicina esteve focado em sistemas especialistas e algoritmos de aprendizado de máquina voltados para a análise de dados clínicos e imagens médicas (Amisha *et al.*, 2019; Jiang *et al.*, 2017; Patel *et al.*, 2009). Contudo, quando se trata de popularidade e usabilidade, não há paralelos com o surgimento de *chatbots* sofisticados como o ChatGPT (OpenAI, EUA), lançado em 2022 e rapidamente alcançando um engajamento massivo — cerca de 100 milhões de usuários em dois meses —, evidenciando não apenas sua facilidade de uso, mas também sua ampla aplicabilidade em diversos campos, incluindo a medicina (Amaral, 2023). Essa popularidade imediata e o interesse crescente nesse tipo de ferramenta deflagraram um expressivo aumento no número de publicações científicas médicas com interesse em IA, alcançando métricas sem precedentes: em 2024 foram encontradas 52.984 publicações com o termo “inteligência artificial” no MEDLINE, o que demonstra forte tendência de alta quando comparado à segunda metade da década de 2010, período no qual a quantidade de citações era significativamente menor, como ilustrado no GRÁFICO 1.

Gráfico 1. Número de publicações por ano (1976-2024) com o termo "Artificial Intelligence", PubMed.



Fonte: elaborado pelo autor a partir de dados disponíveis no MEDLINE/Pubmed

A popularidade dos *chatbots* baseados em modelos de linguagem de larga escala pode ser atribuída a sua capacidade de realizar tarefas complexas de forma acessível e interativa. Na medicina, esses *chatbots* estão sendo explorados para resolver problemas diagnósticos, auxiliar na pesquisa científica e, até mesmo, atuar como ferramenta educacional para estudantes e profissionais. Dentre as várias investigações acerca das potencialidades do uso dos *chatbots* na medicina, destacam-se: (i) o auxílio ao diagnóstico clínico, uma vez que estudos demonstraram que o ChatGPT é capaz de resolver casos clínicos com taxas de precisão diagnóstica comparáveis a médicos humanos em distintos cenários clínicos (Hirosawa *et al.*, 2023; Kanjee; Crowe; Rodman, 2023; Nori *et al.*, 2023); (ii) a demonstração de que a IA pode ser aprovada com alta taxa de acertos em rigorosos testes acadêmicos e de certificação para exercício da medicina, como o USMLE (Kung *et al.*, 2023); (iii) a facilitação da pesquisa científica, isto é, os *chatbots* podem auxiliar na organização e sumarização da literatura científica, de modo que possibilitam otimizar o tempo de pesquisadores (Sallam, 2023).

2.2.2 Aplicações da IA na Educação Médica

O campo da educação médica — termo que contempla um *continuum* de aprendizagem ao longo da vida, que inicia na graduação, passa pela pós-graduação e especialização e vai além, no que se entende como "educação continuada" — demorou a adotar amplamente a IA, ganhando destaque sobretudo nas últimas décadas, devido à aceleração do desenvolvimento tecnológico e à crescente necessidade de métodos educacionais inovadores (Chan; Zary, 2019).

A primeira iniciativa de utilizar recursos de IA para a educação médica foi estabelecida na universidade de Stanford, onde pesquisadores desenvolveram o GUIDON (Clancey, 1983), programa especializado voltado para o ensino da resolução de problemas diagnósticos. Esse *software* foi desenhado para orientar o estudante no diálogo sobre um paciente suspeito de ter uma infecção, utilizando as regras do MYCIN — programa de apoio ao diagnóstico de infecções e prescrição de antibióticos — como material de estudo (Shortliffe, 1976). Dessa forma, ensinava ao estudante dados clínicos e laboratoriais, bem como a forma de utilizar essas informações para diagnosticar o organismo causador da infecção. Essa abordagem inicial abriu caminho para uma série de avanços no uso da inteligência artificial no ensino médico.

Contudo, por décadas o progresso permaneceu limitado devido às restrições tecnológicas e à ausência de disciplinas de tecnologia nos currículos médicos (Chan; Zary, 2019). A partir de meados da década de 1990, com o surgimento de ferramentas de tutoria e simulações baseadas em realidade virtual, como o The Cardiac Tutor, novos programas permitiram que estudantes praticassem habilidades em ambientes digitais e recebessem *feedback* imediato e personalizado (Chan; Zary, 2019; Eliot; Woolf, 1995). A partir de então foi possível notar a preocupação dos programadores em separar o conteúdo a ser ensinado da estratégia de ensino, ou seja, distinguir o "o quê" do "como" ensinar (Eliot; Woolf, 1995).

Com o aumento progressivo dos recursos de IA como ML, DL e redes neurais, os programas de computador usados no ensino foram se tornando mais refinados e capazes de promover aprendizagem adaptativa e *feedback* individualizado, passando a ser mais amplamente utilizados para avaliação dos estudantes (Chan; Zary, 2019; Tolsgaard *et al.*, 2023). Apesar de terem adquirido maior sofisticação tecnológica, até recentemente as capacidades dos sistemas

de IA eram utilizadas de forma pontual e para finalidades específicas, por exemplo: avaliar habilidades em procedimentos psicomotores, simulações conforme protocolos pré-estabelecidos e aprendizagem adaptativa relacionados a situações clínicas bem delimitadas (Chan; Zary, 2019).

Assim, o surgimento de *chatbots* avançados traz o potencial de impactar a educação médica ao introduzir ferramentas interativas e personalizadas para o ensino e a avaliação de estudantes (Gordon *et al*., 2024; Lee, 2023). Ao contrário dos sistemas especializados anteriores, os *chatbots* demonstram capacidades até então indisponíveis, pois permitem explorar tantos temas quantos corpos textuais puderem ser acessados. Esses recursos permitem aos estudantes interagirem com cenários clínicos realistas, conduzirem diagnósticos e justificarem hipóteses em ambientes controlados, enquanto recebem *feedback* em tempo real (Tolsgaard *et al*., 2023).

Porém, desafios persistem na integração de *chatbots* e IA ao ensino médico, o que inclui preocupações éticas, como a presença de vieses e a preservação das competências humanas fundamentais (OMS, 2021). As impressões atuais da literatura sugerem que, para maximizar o impacto dessas ferramentas naturalmente dinâmicas, será necessário desenvolver currículos que integrem a IA à uma formação sólida em ética, com a compreensão de que o aprendizado dos estudantes precisa ser derivado da interação com a IA, e não simplesmente fornecido (ou extraído) pela máquina (Gordon *et al*., 2024; Grunhut; Wyatt; Marques, 2021).

3 PRODUÇÃO BIBLIOGRÁFICA

3.1 Reprodução do artigo desenvolvido e publicado

HOW LARGE LANGUAGE MODELS CHATBOTS CAN FACILITATE THE TEACHING OF CLINICAL REASONING: A SCOPING REVIEW

Guilherme Freitas Bernardo Ferreira – Universidade Professor Edson Antonio Velano
gfreitasbernardo@gmail.com – ORCID 0009-0006-2383-7462

Alexandre Sampaio Moura – Faculdade Santa Casa BH
alexandremoura@faculdadesantacasabh.edu.br – ORCID 0000-0002-4818-5425

Ligia Maria Cayres Ribeiro – Wenckebach Institute (WIOO); Lifelong Learning, Education and Assessment Research Network (LEARN); University Medical Center Groningen, The Netherlands
l.m.cayres.ribeiro@umcg.nl – ORCID 0000-0002-4277-3066

Maria Aparecida Turci – Universidade Professor Edson Antonio Velano
mariaturci@gmail.com – ORCID 0000-0002-4380-4231

Silvia Mamede – Institute of Medical Education Research Rotterdam, Erasmus Medical Center, Erasmus University Rotterdam, The Netherlands
silviamamede@gmail.com – ORCID 0000-0003-1187-2392

3.1.1 Abstract

Introduction: Since the release of large language models (LLM) chatbots, many attempts have emerged to explore their use in patient care and medical education. The use of these as clinical reasoning-supporting tools is being extensively explored, but the extent to which its use is

theory and evidence-informed is yet to be explored. This review aims to map and summarize the state of current research to identify the applicability of LLM chatbots for teaching clinical reasoning during medical training, considering the best evidence available on clinical reasoning development. **Methods:** This scoping review employed Arksey and O'Malley's framework. A systematic and comprehensive search was conducted across PubMed/MEDLINE, Web of Science, and Google Scholar. We included publications that provided insights on how LLM chatbots could help teach clinical reasoning. We adhered to what previous reviews pointed out as evidence-based strategies to teach clinical reasoning effectively. **Results:** The review included 21 publications. All the studies explored ChatGPT (OpenAI); three (14%) additionally explored the use of Bard (Google), two (9,5%) the use of Bing (Microsoft), and one (5%) explored other AI tools. Our findings demonstrated that LLM chatbots can provide coherent differential diagnostic lists, explain their answers in clinical tests, and interact as a tutor in a previously trained simulation scenario. Three articles reported risks of bias and hallucinations when interacting with chatbots. **Discussion:** Our findings suggest that LLM chatbots can support the development of clinical reasoning skills through effective educational strategies. Chatbots' responses can help students build understanding, promote deliberate reflection, foster feedback when practicing with written cases, and adapt the content to the stage of learning. Few studies raised concerns regarding risks and ethical issues. **Conclusion:** This review has demonstrated that LLM chatbots hold significant promise for enhancing clinical reasoning development during medical training. However, addressing inherent limitations, such as the risks of hallucinations and inaccurate explanations, is crucial to maximize their educational potential.

3.1.2 Introduction

Since the release of large language model (LLM) chatbots such as ChatGPT (OpenAI), many attempts have emerged to explore their use in patient care and medical education. The potential use of these chatbots as clinical reasoning supporting tools for physicians has been extensively explored. A similar trend has been observed in medical training, with a growing number of attempts to use these tools to foster students' clinical reasoning (1). While the responsible use of LLMs in clinical practice depends critically on human clinical reasoning expertise,

developing such expertise can also benefit from this technology. Nevertheless, whether and how LLM can help teach clinical reasoning during medical training remains unclear. This paper reports on a review that searched for educational strategies based on LLM chatbots that could be potentially employed in the teaching-learning of clinical reasoning.

The doctor's way of solving a clinical problem has been studied since the 1970s (2). *Clinical reasoning* is a complex set of "*skills, processes, or outcomes wherein clinicians observe, collect and interpret data to diagnose and treat patients."* Therefore, clinical reasoning is essential in a doctor's performance and crucial for diagnostic accuracy and therapeutic success.

Previous studies have shown that expertise in medicine develops through a process where biomedical knowledge gradually becomes encapsulated and integrated with clinical knowledge into illness scripts (4). Initially, students form detailed biomedical networks to explain signs and symptoms, which later transition to form illness scripts. These 'packages' are cognitive schemas that become increasingly refined through experience with clinical problems. They encompass little pathophysiological knowledge but are rich in relevant information about enabling conditions of a given disease and the signs and symptoms with which it presents itself. With that in mind, teachers should be aware that it is impossible to transfer the ability to reason to solve problems linearly. This is so because clinical reasoning relies on an extensive base of acquired knowledge and depends on constructing and activating relevant mental scripts (2, 5-7).

Instead, a more comprehensive approach to clinical teaching should be taken, focusing on fostering the development of large sets of well-organized illness scripts and recognizing the benefits of both analytic and non-analytic reasoning approaches (7). Analytical reasoning involves conscious control and demands effort; it is slow, whereas non-analytical reasoning is fast and effortless, usually based on heuristics and pattern recognition. These two modes of reasoning apply to the diagnosis process (2). Effective teaching strategies should emphasize the importance of using diverse examples to build a robust mental database for students, integrate biomedical and clinical concepts, mimic real-life scenarios, and improve understanding and diagnosis by recognizing underlying similarities across different medical issues (7).

Many strategies have been empirically or systematically developed to teach clinical reasoning, but few have been proven effective (8). There is a lack of evidence supporting the idea that teaching general thinking processes improves performance. Conversely, specific teaching strategies aimed at building knowledge and understanding have been shown to lead to improvements (9). Some of the interventions with proven benefits are strategies that (1) build an understanding of causal mechanisms of diseases, such as self-explanation, (2) foster comparing and contrasting alternative diagnoses such as structured or deliberated reflection while practicing with cases, (3) enable practice with entire cases with provision of feedback, (4) employ retrieval practice, that is, promote effortful recall of information during teaching (5) promote learning by comparing and contrasting discriminating features of different diagnosis, (6) match the student's stage of learning.

The educational challenge of teaching clinical reasoning can benefit from new technologies. In particular, the recent emergence of chatbots based on large language models (LLM chatbots) sparked increased interest in their applications in medical education (10). This interface is built on generative Artificial Intelligence (AI), which can predict the next word in a given sentence and allows easy interaction between the user and the machine as if chatting.

A scoping review conducted by Preiksaits and Rose revealed that AI holds transformative potential for medical education as it offers exciting opportunities (11). The authors have identified that AI could help improve understanding, work as a continuous education and self-directed learning tool, develop personalized learning plans, and provide feedback (11).

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

3.1.3 Methods

Overview

A scoping review of the literature followed the framework proposed by Arksey and O'Malley (12).

Research question

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

Search for research evidence

Searches were made on the following databases: Pubmed (Public MEDLINE), WoS (Web of Science), and Google Scholar. The following keywords were identified and considered relevant in the definition of key concepts: "chatGPT", "chatbot", "large language model", "artificial intelligence", "medical education", "medical school", "medical student", "clinical education", "clinical reasoning", "clinical judgment", "clinical decision-making". The search strings are displayed in Appendix 1. We restricted our search to original articles in English.

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

Study selection

We aimed to select original research articles published in English to investigate the use of LLM chatbots in diagnostic decision-making with primary or secondary outcomes that could be useful in designing or implementing educational activities based on strategies that have proved effective for teaching clinical reasoning. As "effective," we considered educational strategies that have been pointed out by reviews (8,9) as such: 1) promote structured reflection, 2) build understanding through explanations on the reasoning behind solving a clinical vignette/self-explanation, 3) promote comparing/contrasting between differential diagnosis, 4) encourages practice with cases with the provision of feedback. We also included articles that investigated tools that could help reduce teachers' workload, for example, studies in which the primary outcome was to elaborate test items. Preprint articles and perspective articles, as well as publications that debated computational aspects of AI and explored diagnostic accuracy in clinical settings without correlating this ability to the teaching-learning environment, were outside the scope of the review.

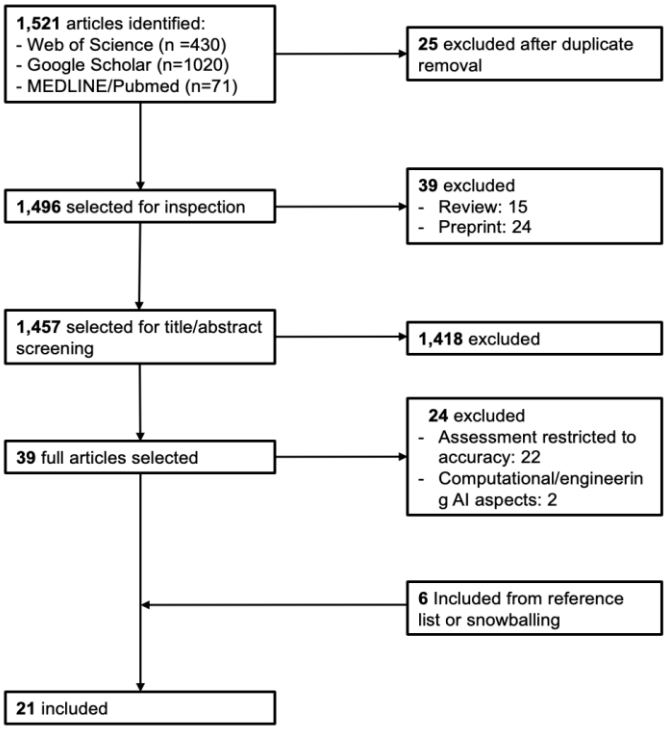
Two independent reserarchers initially selected the articles. After excluding pre-prints, the articles were analyzed according to title and abstract for relevance and were selected after full reading. Any disputes were resolved by agreement between both of them. The following data were charted: Authors, Country, Year Published, Research Method, Outcomes evaluated/questions asked, LLM Interface/Model, Main results, and Relevance for Clinical Reasoning and entered into a 'data charting form.' Then, they were summarized, and the results were reported.

3.1.4 Results

Our search initially identified 1,521 publications. After the exclusion of duplicates, preprints, and perspective articles, we screened the remaining 1,457 according to title and abstracts. We then excluded 1,418 that were considered unrelated to our research. The remaining 39 articles were selected for full reading, and twenty-four were excluded because no correlations with clinical reasoning teaching were found or debated exclusively computational aspects of AI. The

other six articles were included by snowballing or searching the list of references after the concordance of the two authors responsible for the initial selection of the articles. Thus, a total of 21 articles were included in this review. The search and selection process are summarized in Figure 1.

Figure 1. Prisma flow diagram



Source: elaborated by the authors.

Studies Characteristics

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

Five (24%) articles reported chatbot diagnostic accuracy and differential diagnostic lists generation (13-17), ten (48%) focused on the performance in standard tests or exams (14, 18-26), one (5%) measured clinical reasoning skills as the primary outcome (27), two (14%) analyzed chatbot performance on simulation scenarios (one as a tutor and one as a participant) (28,29), and four (19%) evaluated its ability to write clinical exams (30-33). Details of the studies are summarized in Table 1.

Table 1. Summary of findings							
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**
Balas and Ing.	Canada	2023	Quantitative	Providing the most likely and the differential diagnoses for ophthalmologic clinical vignettes	GPT-3	GPT-3 h ad a 90% diagnostic accuracy. The correct diagnosis was included in all differential diagnosis lists generated and in the first median position.	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating and defining factors (similar to a "cued" reflection).
Balasanjeevi and Surapaneni	India	2024	Quantitative and Qualitative	Solving 30 multiple-choice questions extracted from a textbook in Respiratory Medicine	GPT-3.5 GPT-4	Correctness: GPT-3.5: 21/30, GPT-4: 24/30. As a pedagogical tool, ChatGPT4 seemed to showcase the potential to offer deeper explanations and interactive learning experiences.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Bonetti *et al*.	Italy	2023	Quantitative and Qualitative	Solving multiple-choice questions for the Italian Residency Admission National Exam and explaining the answer	GPT-3	GPT 3 answered 87% (122/140) of the questions correctly. Two incorrect answers were due to a logical error (GPT explained the answer correctly but provided the wrong multiple-choice answer) Explanations of the correct answers were all evaluated as appropriate.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.

Authors	Country	Year Published	Research Method	Outcomes evaluated / Questions asked	LLM Interface / Model	Main results	Relevance for Clinical Reasoning
Cai *et al*.	USA	2023	Quantitative and qualitative	Solvig exam questions used by medical residents to prepare for certification in Ophthalmology (Basic and Clinical Sciences)	GPT-3.5, GPT-4, Bing	GPT-4's performance in the test was similar to that of humans. However, it performed less well in questions requiring image interpretation and multiple-step diagnosis. The prevalence of hallucinations for GPT-4 was 18%.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student. However, providing inaccurate explanations may limit the use.
D'Souza *et al*.	India	2023	Qualitative	Solving a hundred psychiatry clinical vignettes from a textbook	GPT-3.5	Test Performance and Preparation (grade A: 61, B: 31, C:8) The authors refined results by categories, which included clinical reasoning and ethical reasoning.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
E *et al*.	Israel	2023	Quantitative and qualitative	Writing a multiple-choice question medical examination	GPT-4	Exam quality, reviewed by specialists, showed that adjustments were needed for 15% of the questions. Only one question (0.5%) out of 210 required replacements due to a completely mistaken answer.	Reduction of teacher workload when preparing clinical reasoning assessments. Interaction for active learning can aid students in refining their learning. Self-learning tool.
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**

Fonseca *et al.*	Portugal	2024	Quantitative	Solving 188 questions from the American Academy of Neurology's Question of the Day app.	GPT-3.5	Score: 85% "AI chatbot provided an adequate explanation for its correct answers in 123 out of 128 cases (96.1%)."	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.
Hirosawa *et al.*	Japan	2023	Quantitative	Solving clinical vignettes in Internal Medicine designed for medical students and junior residents	GPT-3.5	The correct diagnosis was present in 28 out of 30 "top 10 diagnosis" lists provided by GPT and was the top diagnosis in 16. The consistency of the generation of differential diagnosis was 70.5%.	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/defining factors (similar to a "cued" reflection)
Hudon *et al.*	Canada	2024	Quantitative and Qualitative	Crafting Scripts Generation Tests (SCT)	GPT-3.5	"Participants could not identify which SCT was created by ChatGPT from those created by experts in the field." "Most respondents (29/39, 74%) reported the SCTs generated by ChatGPT as caricatural or stereotypical clinical presentations as observed in textbooks."	Reduction of teacher workload when preparing clinical reasoning assessments. Interaction for active learning can aid students in refining their learning. Self-learning tool.
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**
Kanjee *et al.*	USA	2023	Quantitative	Solving cases from NEJM	GPT-4	Diagnostic accuracy was 39%. The mean length of	A generated list of differential diagnoses can help guide the student about diseases on which

				Clinicopathologic conferences		the differential diagnosis list was 9. The correct diagnosis was present in 64% of the generated lists (median position 2.5); the mean quality of the differential diagnosis (accuracy/utility) was 4.2 (in a scale of 0 to 5 proposed by the authors).	they should focus their study of discriminating/defining factors (similar to a "cued" reflection).
Kiyak and Emekli	Turkey	2024	Qualitative	Crafting Scripts Generation Tests (SCT) aimed at undergraduate medical students, with a focus on providing a diagnosis	GPT-4, GPT-4o, Claude 3, Llama 3	"Generated SCT items appear promising, but they have some limitations, such as the absence of a detailed Likert scale description."	Reduction of teacher workload when preparing clinical reasoning assessments.
Koga *et al*.	USA	2023	Quantitative	Solving cases from Mayo Clinic`s Clinicopathologic conferences of neurodegenerative diseases	GPT-3.5, GPT-4, Google Bard	Diagnostic accuracy was 52% for GPT-4, 40% for Bard, and 32% for GPT3.5. The correct diagnoses were included in 84% of the differential diagnosis list generated by ChatGPT-4.	Facilitate Discussions by novice participants. Differential diagnoses and explanations can be helpful in the development of illness scripts.
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**
Kung *et al*.	USA	2023	Quantitative and qualitative	Solving questions from USMLE (excluded questions with images or other visual data)	GPT-3	With indeterminate responses censored/included, ChatGPT accuracy for USMLE Step 1 was 75.0%/45.4%; for Step 2CK, it was 61.5%/	"Partial ability to teach medicine by surfacing novel and nonobvious concepts that may not be in learners' sphere of awareness." Explanations generated by LLM could be used to confront a "self-

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

						"Because of ChatGPT's heralded creativity, it is a potentially inexhaustible source of practice. Whereas standard question banks have vast but finite question pools to choose from, ChatGPT can continuously generate new and unique clinical situations."	
Shea *et al*.	Hong Kong	2023	Quantitative	Solving real clinical cases admitted to a Geriatric ward	GPT-4	The diagnostic accuracy of GPT-4 was 67.% (4/6) The correct diagnosis was present in five out of six lists of Differential Diagnosis generated by GPT-4	It can increase confidence in diagnosis and alert "missing diagnosis." A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/defining factors (similar to a "cued" reflection)
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**
Shieh *et al*.	USA	2024	Quantitative	Solving 109 Step 2 (Clinical Knowledge) from USMLE	GPT -3.5, GPT-4.0	Score: GPT-3.5 47,7%, GPT-4.0 87,2%. ChatGPT 4.0 accurately created a shortlist of differential diagnoses in 74.6% of the 63 case reports.	A generated list of differential diagnoses can help guide the student about diseases on which they should focus their study of discriminating/defining factors (similar to a "cued" reflection)
Strong *et al*.	USA	2023	Quantitative and qualitative	Solving clinical case questions used for clinical reasoning assessment of first- and second-year medical students	GPT-3.5, GPT-4	GPT4 scored higher than students in clinical reasoning skills analysis that involved providing a case summary, a problem	The summaries provided can be used as a cued reflection. AI's capacity to describe its reasoning can help students refine illness scripts and diagnostic schemas.

						list, a diagnostic schema, a differential diagnosis list, and an illness script)	
Wong	USA	2024	Qualitative	Crafting unique clinical cases to be used with first-year graduate medical students	GPT-3.0	"Faculty felt that ChatGPT provided fairly accurate medical statements but could not "clinically reason" or build complexity." "Case creations were too "simplistic" but had the benefit of saving time and providing grammatically correct sentence structure."	Reduction of teacher workload when preparing clinical reasoning assessments.
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**
Xie *et al*.	Australia	2023	Qualitative	Solving simulated clinical cases created by the authors, presented in steps with prompts containing clinical information gathered through the viewpoint of a junior doctor.	GPT-4, Bard, BingAI	Responses were graded for readability using "a combination of the Flesch Reading Ease Score, the Flesch–Kincaid Grade Level, and the Coleman-Liau Index," and also suitability using the DISCERN score (for assessing the responses' quality, relevance, and equitable distribution of information). "When comparing the three LLMs for readability and reliability,	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.

						ChatGPT consistently outperformed, registering the highest Flesch Reading Ease Score (31.2 3.5), Flesch-Kincaid Grade Level (13.5 0.7), Coleman-Lau Index and DISCERN score (62 4.4)"	
Authors	**Country**	**Year Published**	**Research Method**	**Outcomes evaluated / Questions asked**	**LLM Interface / Model**	**Main results**	**Relevance for Clinical Reasoning**
Yiu and Lam	UK	2023	Quantitative/Qualitative	Solving questions from " a mock paper consisting of 300 questions deemed representative of the examination was taken from a popular question bank23 used widely by surgical trainees when preparing for the Royal College of Surgeons membership examination."	GPT-4, Bard	The test performance of GPT was 85,7% without justification and 84,3% with forced justification. A qualitative analysis showed that LLMs might not distinguish between current and outdated guidance in their training datasets, and there were instances where responses displayed clinically inaccurate justifications.	Explanations generated by LLM could be used to confront a "self-explanation" given by the student.

Diagnostic Accuracy and Differential Diagnostic Lists

LLM chatbots showed overall good diagnostic accuracy, at least comparable to experienced physicians in Internal Medicine, Ophthalmology, and Geriatrics. ChatGPT's latest interface performed consistently better than its previous versions and was overall more accurate than other chatbots (34). Interestingly, diagnostic performance was lower when dealing with complex diagnostic challenges discussed at clinicopathological conferences (15, 34).

The LLM chatbots showed a remarkable ability to generate differential diagnostic lists, which were usually comprehensive and included up to ten items. However, this number varied according to the prompting used by researchers. These lists included a high rate of correct diagnoses and were considered accurate and coherent with the presented clinical cases. Of the LLM chatbots investigated, GPT-4 was considered more accurate than GPT-3.5 and Google Bard (34). Three (60%) articles reported risks of bias and "hallucinations" (i.e., when LLM chatbots generate factually incorrect answers) when interacting with chatbots (13-15).

Test Performance

Ten (48%) studies investigated how LLM chatbots would perform in standard tests in different contexts. Three (50%) reported that LLM chatbots achieved passing rates in board certification exams, including USMLE (19, 21, 22). One has found that GPT-3 would be admitted to any residency program in Italy since it raked in the top 98,8th percentile among more than 15,000 candidates (18).

The tests varied significantly regarding the types of questions (e.g., multiple-choice vs. open-ended), reasoning levels (single-step vs. multiple-steps), and fields of knowledge. GPT-4 performed better than its previous versions and other LLM chatbots.

Noteworthy, LLM chatbots usually provide explanations for the questions solved. Investigators scrutinized these explanations and often considered them coherent and appropriate (18, 21, 23, 24). For example, in the study by Kung *et al.*, ChatGPT correctly noted that *"adrenal hypercortisolism increased bone osteoclast activity, which increased*

calcium resorption decreased bone mineral density, which increased fracture risk" (21). However, some concerns were raised (19, 22, 26), such as those presented by Cai *et al*., where ChatGPT provided "an accurate description of trabeculectomy but neglected *"to mention that an aqueous shunt is the preferred procedure in this specific scenario, which could impact clinical decision-making*" (19). In their study, hallucination was also frequent (18%). Another study by You & Lam noted that *"LLMs achieve high concordance and provide insightful responses to test questions,"* but *"inappropriate or inaccurate decision-making, incomplete appreciation of nuanced clinical scenarios and utilization of out-of-date guidance was, however, noted*."

Clinical Reasoning Skills

Strong *et al*. compared ChatGPT's performance (versions 3.5 and 4.0) of medical students in a clinical reasoning test that analyzed a clinical vignette and composed a *"summary of this case in 200 words or less, including a statement as to the most likely diagnosis*". The authors reported that the latest version (GPT-4) scored significantly better than GPT-3.5 and similar to graduate medical students. Passing rates were similar between GPT-4 and students (93% vs 85%). Both were significantly higher than GPT3.5 (43%). The authors showed that GPT-4 could provide better problem lists, and other skills (diagnostic schema, differential diagnosis, illness scripts, case summary) matched those presented by the students in this test.

Simulation Scenarios

Two articles evaluated the LLM chatbot's abilities in designing simulated clinical cases. One (28) investigated whether ChatGPT could act as a tutor to provide instructions in a written case simulation, give feedback, and adapt case progression according to the user's responses. Although it suggests it can be a promising tool in the simulation field, there were some limitations regarding inaccuracy, technical errors, and risk of bias. The authors mentioned that *"ChatGPT occasionally provided feedback on the appropriateness of a decision and then progressed the simulation as if the correct decision had been made"* and considered some of the feedback provided weak and unclear.

Another study (29) explored ChatGPT, Bard, and Bing's performances as respondents to increasingly complex written clinical scenarios. The authors showed that of the three models, ChatGPT is currently the most reliable model regarding comprehensibility and alignment with clinical guidelines. They also alerted for *"misleading or inaccurate responses due to biases in the training data or misconceptions of intricate medical concepts.*"

Writing of Clinical Exams

Four studies explored the potential of GPT-4 to create medical examinations (30-33). After being tutored with an existing model, the machine made a 210 multiple-choice medical examination. Only 0,5% of the questions were labeled entirely inaccurate by the investigators, and 15% required some revision. Two authors explored ChatGPT`s ability to craft Scripts Concordance Tests and demonstrated that although promising (33), severe limitations existed, such as caricatural or stereotypical clinical presentations (32). Hudon *et al.* noted that users couldn't accurately tell if a human or an LLM chatbot created a test.

3.1.5 Discussion

This review aimed to understand better how LLM chatbots could help teach clinical reasoning during medical training, considering the best evidence available on clinical reasoning development. Our findings suggest that LLM chatbots have overall good diagnostic accuracy when dealing with written clinical vignettes, provide coherent differential diagnostic lists (13-15, 17, 25, 34), provide explanations to their answers in clinical tests (18-27), match clinical reasoning skills similar to those students that received specific training (27), and can interact as a tutor in previously trained simulation scenarios (28, 29).

Based on findings of recent systematic reviews (8, 9) that pointed out evidenced-based strategies to teach clinical reasoning, we aimed to correlate these strategies with the skills demonstrated by LLM chatbots.

Strategies aimed at building understanding

The explanations provided by chatbots could be confronted with a self-explanation given by the student during practice with clinical cases (9). Interestingly, LLM chatbots provided explanations that contained nonobvious insights, requiring deduction or knowledge external to the question input (21). Also, Strong *et al*. (27) investigated ChatGPT's clinical reasoning ability and demonstrated that it is at least comparable to that of students with formal training.

The ability shown by chatbots to summarize clinical data and describe the reasoning behind the provided responses can act as a source of feedback, helping students reflect on and refine their explanations for a problem. This may contribute to restructuring knowledge of causal mechanisms underlying diseases and refine illness scripts according to their level of expertise. It has been demonstrated that knowledge of causal mechanisms acts as a "glue" that helps link clinical findings together, thereby facilitating the recognition of diseases (35). This can ultimately increase diagnostic accuracy and provide better results for patients in the future.

However, problems inherent to artificial intelligence, such as hallucinations and providing inaccurate explanations, may limit its use, at least in the models currently available (13-15).

Promote reflection

Structured reflection during practice with clinical cases has proven beneficial for building students' knowledge and reinforcing illness scripts and diagnostic schemas, thereby improving students' diagnostic performance (8,9,36). By generating a broad, accurate differential diagnosis list, LLM chatbots can potentially help guide the students' reflection. The chatbot lists can help students identify which diseases they should consider when reflecting upon a particular clinical case and help focus their study on discriminating/defining factors, similar to what happened in studies that used a "cued" reflection (37). We have found that the lists provided by chatbots are sufficiently accurate and reliable. Still, they should only be used when guided by specialists or teachers due to the risk of bias and issues mentioned above.

Practice with entire cases with the provision of feedback

Practice with a large sample of clinical problems is considered critical for developing clinical reasoning (7). Moreover, research in many domains has demonstrated that practice with problems associated with corrective feedback is the primary mechanism for developing expertise (38). It is likely, therefore, that providing students with feedback when they practice with clinical cases would be beneficial. Scherr *et al*. (25) demonstrated that GPT-3.5 could act as a conductor in two critical care common scenarios with post-scenario feedback. Although the benefits of simulation-based learning are beyond the scope of this review, we believe that by doing so, LLM chatbots can help self-directed learning. We highlight that the risks of hallucinations and the limited number of studies exploring this type of interaction with LLM chatbots impose significant limitations.

Adaptation to the stage of learning

Teaching should be tailored appropriately to each stage of learning (9). Koga (34) suggests that LLM chatbots can facilitate discussions by novice participants when participating in complex clinicopathologic discussions. Another study (22) argued that LLM chatbots could provide "nonobvious concepts that may not be in learners' sphere of awareness." Also, LLM chatbot adaptative tutoring was discussed above. We believe these characteristics give chatbots the potential to help foster active learning and consolidate illness scripts and mental schemas.

Concerns regarding AI and clinical reasoning education

A few studies mentioned concerns about ethical aspects, risks for students and patients, and the exacerbation of inequalities related to access to available technologies (20-22, 29). One study noted that due to the nature of chatbots, meaning that they extract information from web-based sources, their generated clinical scenarios and responses "may reflect societal and systemic biases that already exist in medical education" (28). In addition, some studies call attention to privacy and data security threats, especially when considering learning in real clinical settings (10, 39, 40). Lastly, most published studies on AI and clinical reasoning aimed to report the machine's capabilities to pass a standard test or establish correct diagnoses. Only a few studies sought to delve into the interaction

between students and artificial intelligence to understand how these can enhance clinical reasoning skills among students. This would be critical to guide the incorporation of the tools in education. Nevertheless, it seems clear, particularly considering the limitations of the present AI models, that teachers should use technology as an ally in clinical reasoning education, not as a substitute.

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

3.1.6 Conclusion

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

3.1.7 References

1. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, *et al*. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Medical Teacher. 2024 Feb 29;1–25.
2. Mamede S. What does research on clinical reasoning have to say to clinical teachers? Sci Med. 2020 Jul 15;30:1–8.
3. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, *et al*. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Academic Medicine. 2019 Jun;94(6):902–12.
4. Schmidt HG, Rikers RMJP. How expertise develops in medicine: knowledge

encapsulation and illness script formation. Med Educ. 2007 Nov 14;0(0):071116225013002-???
5. Bowen JL. Educational Strategies to Promote Clinical Diagnostic Reasoning. n engl j med. 2006;
6. Cutrer WB, Sullivan WM, Fleming AE. Educational Strategies for Improving Clinical Reasoning. Current Problems in Pediatric and Adolescent Health Care. 2013 Oct;43(9):248–57.
7. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005 Jan;39(1):98–106.
8. Prakash S, Sladek RM, Schuwirth L. Interventions to improve diagnostic decision making: A systematic review and meta-analysis on reflective strategies. Medical Teacher. 2019 May 4;41(5):517–24.
9. Cooper N, Bartlett M, Gay S, Hammond A, Lillicrap M, Matthan J, *et al*. Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. Medical Teacher. 2021 Feb 1;43(2):152–9.
10. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anatomical Sciences Education. 2023;
11. Preiksaitis C, Rose C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. JMIR Medical Education. 2023;9(1):e48785.
12. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005 Feb;8(1):19–32.
13. Balas M, Ing EB. Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. JFO Open Ophthalmology. 2023 Mar;1:100005.
14. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. IJERPH. 2023 Feb 15;20(4):3378.
15. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023 Jul 3;330(1):78.
16. Koga S. The Potential of ChatGPT in Medical Education: Focusing on USMLE Preparation. Ann Biomed Eng. 2023 Oct;51(10):2123–4.
17. Shea YF, Lee CMY, Ip WCT, Luk DWA, Wong SSW. Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis. JAMA Netw Open. 2023 Aug 14;6(8):e2325000.
18. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? Ann Biomed Eng. 2023 Jul 25;
19. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, *et al*. Performance of Generative Large Language Models on Ophthalmology Board–Style Questions. American Journal of Ophthalmology. 2023 Oct;254:141–9.
20. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. 2023 Nov;89:103770.
21. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Dagan A, editor. PLOS Digit Health. 2023 Feb 9;2(2):e0000198.
22. Yiu A, Lam K. Performance of large language models at the MRCS Part A: a

tool for medical education? Ann R Coll Surg Engl. 2023 Dec 1;
23. Balasanjeevi G, Surapaneni KM. Comparison of ChatGPT version 3.5 & 4 for utility in respiratory medicine education using clinical case scenarios. Respiratory Medicine and Research. 2024 Jun;85:101091.
24. Fonseca Â, Ferreira A, Ribeiro L, Moreira S, Duque C. Embracing the future—is artificial intelligence already better? A comparative study of artificial intelligence performance in diagnostic accuracy and decision-making. Euro J of Neurology. 2024 Apr;31(4):e16195.
25. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep. 2024 Apr 23;14(1):9330.
26. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, Pérez-Sancristóbal I, Pato-Cour E, Plasencia-Rodríguez C, *et al*. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep. 2023 Dec 13;13(1):22129.
27. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, *et al*. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA Intern Med. 2023 Sep 1;183(9):1028–30.
28. 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:e49877.
29. 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. 2023 Aug 21;ans.18666.
30. Wong K, Fayngersh A, Traba C, Cennimo D, Kothari N, Chen S. Using ChatGPT in the Development of Clinical Reasoning Cases: A Qualitative Study. Cureus [Internet]. 2024 May 31 [cited 2024 Jul 23]; Available from: https://www.cureus.com/articles/253053-using-chatgpt-in-the-development-of-clinical-reasoning-cases-a-qualitative-study
31. E K, S P, R G, R KL, A B, M G, *et al*. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ. 2023 Oct 17;23(1):772.
32. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in Psychiatry to Design Script Concordance Tests in Undergraduate Medical Education: Mixed Methods Study. JMIR Med Educ. 2024 Apr 4;10:e54067–e54067.
33. Kıyak YS, Emekli E. A Prompt for Generating Script Concordance Test Using ChatGPT, Claude, and Llama Large Language Model Chatbots. 2024;
34. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: CHATGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathology. 2023 Aug 8;e13207.
35. Woods NN, Neville AJ, Levinson AJ, Howey EHA, Oczkowski WJ, Norman GR. The Value of Basic Science in Clinical Diagnosis: Academic Medicine. 2006 Oct;81(Suppl):S124–7.
36. Mamede S, Schmidt HG. Deliberate reflection and clinical reasoning: Founding ideas and empirical findings. Medical Education. 2023 Jan;57(1):76–85.
37. Ibiapina C, Mamede S, Moura A, Elói-Santos S, Van Gog T. Effects of free, cued and modeled reflection on medical students' diagnostic competence. Med Educ. 2014 Aug;48(8):796–805.
38. Ericsson KA. Deliberate Practice and the Acquisition and Maintenance of

Expert Performance in Medicine and Related Domains: Academic Medicine. 2004 Oct;79(Supplement):S70–81.
39. Grunhut J, Wyatt AT, Marques O. Educating Future Physicians in Artificial Intelligence (AI): An Integrative Review and Proposed Changes. Journal of Medical Education and Curricular Development. 2021 Jan;8:238212052110368.
40. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ. 2022 Nov 9;22(1):772.

CONTRIBUTIONS

GFBF and ASM searched for and selected the included articles, extracted the primary information contained in the selected articles, and wrote the manuscript. LMCR and SM guided the theoretical foundation for the research and reviewed the manuscript. MAT designed the search and selection method and reviewed the text. All authors participated in the analysis and interpretation of the results, discussed the relevance of the data, and established correlations with the teaching of clinical reasoning.

DISCLOSURE STATEMENT

The authors report no conflicts of interest and are alone responsible for the article's content and writing.

FUNDING

This study did not receive funding from any source.

ETHICAL APPROVAL

As this study reviews previously published studies, institutional ethics committee approval is unnecessary.

APPENDIX 1

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

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

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

4 PRODUÇÃO TÉCNICA

4.1 Guia – “6 dicas práticas para a usar o ChatGPT para facilitar o ensino do raciocínio clínico”

O Guia "6 dicas práticas para a usar o ChatGPT para facilitar o ensino do raciocínio clínico” é um produto bibliográfico voltado para educadores, com o objetivo de fornecer sugestões baseadas em evidências para estimular a utilização desse recurso educacional. Estruturado de forma acessível, o material reúne dicas claras e sugestões simplificadas de *prompts* para facilitar a interação com o ChatGPT. A compreensão de conceitos técnicos é abordada por meio de um glossário de termos.

4.1.1 Dicas

1. Construa listas de diagnósticos diferenciais

Solicite listas de diagnósticos diferenciais detalhadas para casos clínicos simulados.

Dica prática: peça que o ChatGPT justifique cada diagnóstico com base nos achados apresentados. Compare as explicações fornecidas pela IA com as autoexplicações geradas pelos alunos. Enfatize a geração de conhecimento novo e a ampliação do repertório diagnóstico dos estudantes.

***Prompt*:** “Elabore uma lista de possíveis diagnósticos diferenciais para o caso clínico a seguir: [copie e cole a vinheta clínica]. Para cada diagnóstico fornecido, explicite os dados clínicos que o justificam e explique.”

2. Estimule a reflexão estruturada

Utilize as explicações do ChatGPT para guiar reflexões sobre os casos.

Dica prática: compare a explicação gerada pelo *chatbot* com a análise do estudante, destacando lacunas e *insights* adicionais que promovam o aprendizado. Enfatize o processo reflexivo, ajudando os estudantes a consolidarem o conhecimento e ajustarem seus raciocínios.

***Prompt*:** "Com base no caso clínico a seguir, forneça uma explicação detalhada sobre o raciocínio que levou ao diagnóstico final. Depois disso, avalie se há outros diagnósticos diferenciais que poderiam ser considerados e quais informações adicionais seriam necessárias para descartá-los. Caso clínico: [copie e cole o caso clínico]."

3. Elabore simulações e *feedback*

Use o ChatGPT para criar cenários clínicos simulados, ajustando as etapas conforme as decisões dos alunos.

Dica prática: integre o *chatbot* como tutor, oferecendo *feedback* detalhado após cada etapa da simulação para reforçar conceitos e corrigir erros.

***Prompt*:** "Simule um caso clínico progressivo para o seguinte cenário clínico [você pode inserir o caso clínico detalhado ou elaborar a partir da interação com o ChatGPT]. Apresente informações adicionais a cada interação e forneça *feedback* imediato sobre as decisões tomadas."

4. Adapte ao nível de aprendizado

Ajuste o uso do ChatGPT ao estágio de formação dos alunos, promovendo discussões que fortaleçam os esquemas mentais e *scripts* de doenças.

Dica prática: utilize *chatbots* para facilitar a discussão de temas complexos e fornecer conceitos não óbvios, de acordo com o nível de aprendizagem.

Prompt: "Demonstre como o sinal ou sintoma [você pode descrever qual sintoma e doença quer explicar, ou utilizar de algum caso clínico que esteja sendo abordado na mesma interação] se relaciona à fisiopatologia da doença."

5. Construa testes e avalie conhecimentos

Utilize o ChatGPT para criar exames e questões que abordem raciocínio clínico.

Dica prática: combine perguntas para avaliar tanto o conhecimento factual quanto o processo de pensamento. Se preferir, pode incluir no *prompt* o comando "faça as perguntas uma de cada vez".

***Prompt*:** "Crie um conjunto de perguntas [inserir número de perguntas de múltipla escolha e questões abertas] sobre o caso clínico a seguir: [inserir caso clínico]. Use perguntas que avaliem o raciocínio clínico e os conhecimentos diagnósticos do estudante. Inclua as respostas corretas e explicações detalhadas para cada questão."

6. Esteja atento aos riscos e limitações

Explique aos alunos os riscos de "alucinações" (respostas imprecisas) e vieses dos modelos de IA.

Dica prática: oriente sempre o uso supervisionado da tecnologia, reforçando que o ChatGPT é um aliado no aprendizado, não um substituto para a análise crítica. O uso dessa ferramenta no ensino clínico deve ser supervisionado para maximizar o aprendizado, garantindo segurança e precisão. A aprendizagem deve ser derivada da interação com a IA, e não meramente substituída por ela.

4.1.2 Glossário de Termos em IA

- **Alucinações em IA**

As alucinações são dados inventados pela IA que incluem referências ou fontes fabricadas, imprecisas ou desalinhadas, as quais são apresentadas como verdade no texto gerado.

- ***Chatbot***

Chatbot é um programa de computador projetado para simular uma conversa com usuários humanos, tanto por meio de mensagens de texto quanto por voz. Os *chatbots* utilizam técnicas de inteligência artificial, como processamento de linguagem natural, para entender e responder perguntas, realizar tarefas e fornecer informações.

- **ChatGPT**

O ChatGPT é um modelo de linguagem baseado em inteligência artificial, desenvolvido pela empresa OpenAI, capaz de gerar respostas em linguagem natural a partir de comandos textuais. Esse modelo utiliza redes neurais profundas e aprendizado supervisionado para fornecer respostas coerentes e contextualizadas.

- ***Deep learning* (aprendizado profundo)**

O aprendizado profundo é um subconjunto do aprendizado de máquina, a qual faz uso de algoritmos para processar múltiplas camadas de informação para modelar relações entre dados.

- **Gemini**

Gemini é o assistente de IA desenvolvido pelo Google. Ele integra recursos de processamento de linguagem natural, ajudando usuários com respostas automatizadas e tarefas cotidianas.

- **DeepSeek (DeepSeek-R1)**

A DeepSeek é uma empresa *startup* chinesa desenvolvedora de *softwares* com foco em inteligência artificial. Em janeiro de 2025, o lançamento do *chatbot* da empresa (DeepSeek-R1) foi noticiado por ter causado forte queda dos valores de ações de companhias de tecnologia e inteligência artificial, devido ao baixo custo operacional relatado pela empresa.

- **IA generativa**

A IA generativa é um tipo de tecnologia de inteligência artificial que se baseia em modelos de *deep learning* treinados por meio de grandes conjuntos de dados para criar e gerar novos conteúdos.

- **Inteligência Artificial**

Inteligência artificial (IA) é uma disciplina da ciência e tecnologia que permite a computadores e programas realizarem tarefas que tradicionalmente exigem inteligência humana. Ela abrange tecnologias como sistemas especialistas, visão computacional, robótica e aprendizado de máquina. A IA pode executar funções semelhantes às humanas, aprender com a experiência, adaptar-se a novas situações e otimizar seu desempenho em tarefas específicas.

- ***Machine Learning* (aprendizado de máquina)**

Machine learning é um subconjunto da IA no qual utiliza-se um modelo de computador para resolver problemas usando teorias estatísticas ou identificando padrões específicos nos dados. Envolve algoritmos de programação para tomar decisões ou fazer previsões com base em dados.

- **Modelos de Linguagem de Larga Escala**

Modelos de linguagem de larga escala são modelos de aprendizado de máquina, geralmente redes neurais profundas, que processam, interpretam e geram textos prevendo a próxima palavra ou sequência de palavras com base em grandes volumes de dados textuais.

- **Processamento de Linguagem Natural**

O processamento de linguagem natural é um processo de *software* conduzido por inteligência artificial que interpreta e produz linguagem natural escrita ou falada a partir de dados estruturados e não estruturados. Ela ajuda os computadores a fornecer respostas aos usuários em linguagem humana.

- ***Prompt***

Prompt refere-se à instrução, comando ou entrada textual fornecida a um programa de computador para gerar uma resposta. O *prompt* pode variar em termos de complexidade e formato, desde perguntas diretas até comandos mais elaborados, influenciando a resposta gerada pela máquina.

- **Viés em IA**

O viés em inteligência artificial refere-se à ocorrência de resultados tendenciosos que podem estar relacionados a vieses humanos ou não, que distorcem os dados de treinamento originais ou o algoritmo de IA levando a resultados deturpados e potencialmente prejudiciais.

5 CONSIDERAÇÕES FINAIS

Os resultados evidenciam desafios importantes que precisam ser enfrentados para maximizar o impacto da inteligência artificial no ensino do raciocínio clínico. Primeiro, a velocidade de atualização e expansão das tecnologias de inteligência artificial representa um desafio contínuo para a literatura acadêmica. Os estudos sobre ferramentas como o ChatGPT são rapidamente superados pela evolução tecnológica, o que torna difícil manter a relevância e a atualidade das recomendações. Isso reforça a necessidade de práticas educacionais adaptáveis e dinâmicas, que incorporem novas funcionalidades. Além disso, os riscos de vieses, alucinações e conflitos éticos ressaltam a importância de desenvolver currículos que combinem competências tecnológicas com habilidades humanas fundamentais, sempre com atenção ao fato de que o foco na geração de conhecimento é elemento indispensável para o desenvolvimento do raciocínio clínico dos estudantes.

Em segundo lugar, para a revisão bibliográfica foi necessário refletir sobre os dados dos artigos revisados para propor relações com estratégias de ensino do raciocínio clínico, o que, muitas vezes, gerou extrapolações e inferências a partir dos resultados de cada publicação. Embora isso tenha permitido explorar possibilidades, também destacou a ausência de estudos robustos que integrem diretamente as capacidades dos *chatbots* com estratégias pedagógicas baseadas em evidências. Isso aponta para um campo amplo de investigação, que poderá aprofundar o uso dessas ferramentas no contexto do ensino do raciocínio clínico de maneira mais estruturada e empiricamente fundamentada. Por exemplo, a revisão de escopo aponta o potencial dos *chatbots* em fornecer listas diferenciais de diagnóstico coerentes, explicações detalhadas e interatividade no ensino simulado.

Portanto, este estudo demonstra o papel promissor dos *chatbots* no ensino do raciocínio clínico e oferece uma perspectiva sobre a necessidade de uma integração cautelosa e colaborativa entre educadores, pesquisadores em educação e profissionais de saúde para a assimilação dessa tecnologia de maneira proveitosa. Dessa forma, será possível explorar todo o seu potencial, garantindo que o foco permaneça no aprimoramento do aprendizado e na construção do conhecimento dos estudantes.

REFERÊNCIAS

AMARAL, O. O ChatGPT e os limites da inteligência artificial: novos modelos de inteligência artificial colocam a humanidade em questão. **Revista Piauí**. São Paulo, n. 199, 2023.

AMISHA *et al.* Overview of artificial intelligence in medicine. **Journal of Family Medicine and Primary Care**. Mumbai, v. 8, n. 7, p. 2328, 2019.

BOWEN, J. L. Educational Strategies to Promote Clinical Diagnostic Reasoning. **New England Journal of Medicine**. Boston, v. 355, p. 2217-2225, 2006.

CHAMBERLAND, M. *et al*. The influence of medical students' self-explanations on diagnostic performance. **Medical education**. Oxford, v. 45, n. 7, p. 688-695, jul. 2011.

CHAN, K. S.; ZARY, N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. **JMIR Medical Education**. Toronto, v. 5, n. 1, p. e13930, 15 jun. 2019.

CHARNESS, N. The impact of chess research on cognitive science. **Psychological Research**. Berlin, v. 54, n. 1, p. 4-9, mar. 1992.

CIVANER, M. M. *et al.* Artificial intelligence in medical education: a cross-sectional needs assessment. **BMC Medical Education**. Londres, v. 22, n. 1, p. 772, 9 nov. 2022.

CLANCEY, W. J. GUIDON. Technical Report# 9. São Francisco, 1983.

COOPER, N. *et al.* Consensus statement on the content of clinical reasoning curricula in undergraduate medical education. **Medical Teacher**, Londres v. 43, n. 2, p. 152-159, 1 fev. 2021.

CUSTERS, E. J.; REGEHR, G.; NORMAN, G. R. Mental representations of medical diagnostic knowledge: a review. **Academic Medicine**. Filadélfia, v. 71, n. 10, 1996.

DANIEL, M. *et al*. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. **Academic Medicine**, Filadélfia, v. 94, n. 6, p. 902-912, jun. 2019.

ELIOT, C.; WOOLF, B. P. An adaptive Student Centered Curriculum for an intelligent training system. **User Modelling and User-Adapted Interaction**. Boston, v. 5, n. 1, p. 67-86, 1995.

EVA, K. W. What every teacher needs to know about clinical reasoning. **Medical Education**, Oxford, v. 39, n. 1, p. 98-106, jan. 2005.

GORDON, M. *et al.* A scoping review of artificial intelligence in medical education: BEME Guide Nº 84. **Medical Teacher**. Londres, p. 1-25, 29 fev. 2024.

GRUNHUT, J.; WYATT, A. T.; MARQUES, O. Educating Future Physicians in Artificial Intelligence (AI): An Integrative Review and Proposed Changes. **Journal of Medical Education and Curricular Development**, Auckland, v. 8, p. 238212052110368, jan. 2021.

HIROSAWA, T. *et al*. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. **International Journal of Environmental Research and Public Health**, Basileia, v. 20, n. 4, p. 3378, 15 fev. 2023.

JIANG, F. *et al*. Artificial intelligence in healthcare: past, present and future. **Stroke and Vascular Neurology**, Londres, v. 2, n. 4, p. 230-243, dez. 2017.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, Nova Iorque, v. 349, n. 6245, p. aaa8415, 17 jul. 2015.

KAHNEMAN, D. **Fast and slow thinking**. Nova Iorque: Allen Lane and Penguin Books, 2011.

KANJEE, Z.; CROWE, B.; RODMAN, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. **JAMA**, Chicago, v. 330, n. 1, p. 78, 3 jul. 2023.

KUNG, T. H. *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. **PLOS Digital Health**, São Fransico, v. 2, n. 2, p. e0000198, 9 fev. 2023.

LEBEDEV, M. A.; NICOLELIS, M. A. L. Brain-machine interfaces: past, present and future. **Trends in Neurosciences**, Oxford, v. 29, n. 9, p. 536-546, set. 2006.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**. Londres, v. 521, n. 7553, p. 436-444, 27 maio 2015.

LEE, H. The rise of ChatGPT: Exploring its potential in medical education. **Anatomical Sciences Education**, Hoboken, v. 17, n. 9, 2023.

LEE, J. *et al*. Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. **Academic Medicine**, Filadélfia, v. 96, n. 11S, p. S62-S70, nov. 2021.

MAMEDE, S. *et al*. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. **Medical Education**, Oxford, v. 46, n. 5, p. 464-472, maio 2012.

MAMEDE, S. What does research on clinical reasoning have to say to clinical teachers? **Scientia Medica**, Porto Alegre, v. 30, p. 1-8, 15 jul. 2020.

MAMEDE, S.; SCHMIDT, H. G. Deliberate reflection and clinical reasoning: Founding ideas and empirical findings. **Medical Education**, Oxford, v. 57, n. 1, p. 76-85, jan. 2023.

MAMEDE, S.; SCHMIDT, H. G.; PENAFORTE, J. C. Effects of reflective practice on the accuracy of medical diagnoses. **Medical Education**, Oxford, v. 42, n. 5, p. 468-475, maio 2008.

MCCARTHY, J. *et al.* A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. **AI Magazine**, La Canada, v. 27, n. 4, p. 12, 1956.

NORI, H. *et al*. Capabilities of GPT-4 on Medical Challenge Problems. **arXiv**, Ithaca, 12 abr. 2023. Disponível em: <http://arxiv.org/abs/2303.13375>. Acesso em: 1 mar. 2024.

NORMAN, G. Research in clinical reasoning: past history and current trends. **Medical Education**, Oxford v. 39, n. 4, p. 418-427, abr. 2005.

OLIVEIRA, J. D. C. V. *et al*. Ensino do raciocínio clínico orientado pela Teoria dos Scripts de Doenças. **Arquivos Brasileiros de Cardiologia**, São Paulo, v. 119, n. 5, p. 14-21, 9 nov. 2022.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Ethics and Governance of Artificial Intelligence for Health WHO Guidance**. Genebra: Organização Mundial da Saúde, 2021.

PARK, S. H. *et al*. Artificial Intelligence-related Literature in Transplantation: A Practical Guide. **Transplantation**, Londres, v. 105, n. 4, p. 704-708, abr. 2021.

PATEL, V. L. *et al*. The coming of age of artificial intelligence in medicine. **Artificial Intelligence in Medicine**, Amsterdã, v. 46, n. 1, p. 5-17, maio 2009.

PRAKASH, S.; SLADEK, R. M.; SCHUWIRTH, L. Interventions to improve diagnostic decision making: A systematic review and meta-analysis on reflective strategies. **Medical Teacher**, Londres, v. 41, n. 5, p. 517-524, 4 maio 2019.

SALLAM, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. **Healthcare**, Basileia, v. 11, n. 6, p. 887, 19 mar. 2023.

SCHMIDT, H. G.; MAMEDE, S. How to improve the teaching of clinical reasoning: a narrative review and a proposal. **Medical Education**, Oxford v. 49, n. 10, p. 961-973, out. 2015.

SCHMIDT, H. G.; RIKERS, R. M. J. P. How expertise develops in medicine: knowledge encapsulation and illness script formation. **Medical Education**, Oxford, 14 nov. 2007.

SHAH, N. H.; ENTWISTLE, D.; PFEFFER, M. A. Creation and Adoption of Large Language Models in Medicine. **Journal of the American Medical Association**, Chicago v. 330, n. 9, p. 866, 5 set. 2023.

SHEIKHALISHAHI, S. *et al*. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. **JMIR Medical Informatics**. Toronto. v. 7, n. 2, p. e12239, 27 abr. 2019.

SHERBINO, J. *et al*. Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: acontrolled trial. **CJEM**, Ottawa, v. 16, n. 1, p. 34-40, jan. 2014.

SHORTLIFFE, E. Computer-based medical consultations: MYCIN. **Artificial Intelligence – AI**, v. 388, 1 out. 1976.

SIMON, H.; CHASE, W. Skill in Chess. Em: LEVY, D. (Ed.). **Computer Chess**

Compendium. Nova Iorque, NY: Springer New York, 1988. p. 175-188.

TOLSGAARD, M. G. *et al*. The fundamentals of Artificial Intelligence in medical education research: AMEE Guide Nº 156. **Medical Teacher**, Londres, v. 45, n. 6, p. 565-573, 3 jun. 2023.

TURING, A. M. I. Computing Machinery and Intelligence. **Mind**, Oxford. v. LIX, n. 236, p. 433-460, 1 out. 1950.

APÊNDICE

UNIVERSIDADE PROFESSOR EDSON ANTÔNIO VELANO

MESTRADO PROFISSIONAL ENSINO EM SAÚDE

RELATÓRIO DE DESCRIÇÃO DE PRODUTO TÉCNICO/TECNOLÓGICO

PRODUTO: 6 dicas para usar o ChatGPT para facilitar o ensino do raciocínio clínico

DESENVOLVEDOR(ES): Guilherme Freitas Bernardo Ferreira, Maria Aparecida Turci, Luan Marcus Maminhak Morais

ADERÊNCIA

O produto está relacionado à linha de pesquisa do Ensino do raciocínio clínico, e foi elaborado para servir como um pequeno guia para educadores utilizaram-se do ChatGPT para auxiliar os estudantes a elaborar e consolidar *scripts* de doenças e, dessa forma, facilitar o desenvolvimento do raciocínio clínico. O guia tem como referencial teórico a revisão de escopo executada neste trabalho, e é acompanhado de um glossário dos principais termos relacionados às áreas de inteligência artificial e raciocínio clínico.

Dissertação: REVISÃO DE ESCOPO SOBRE O USO DA INTELIGÊNCIA ARTIFICIAL GENERATIVA NO DESENVOLVIMENTO DO RACIOCÍNIO CLÍNICO E A CONFECÇÃO DE GUIA APLICADO AO ENSINO MÉDICO

Linha de Pesquisa: Raciocínio Clínico.

IMPACTO

O guia foi desenvolvido a partir da necessidade de sistematizar os principais achados da revisão de escopo, com o objetivo de fornecer recomendações baseadas em evidências para o educador que tenha interesse em incorporar os *chatbots* no ensino do raciocínio clínico. O produto estará disponível no repositório da UNIFENAS para acesso livre por

docentes de medicina dos cursos brasileiros, podendo também ser consumido pelos demais países de Língua Portuguesa. Além disso, será incorporado material de ensino clínico do Curso de Medicina da UNIFENAS, em Belo Horizonte e Alfenas, podendo se configurar como ferramenta inovadora do ensino clínico e consonante com os anseios da atual geração de estudantes.

APLICABILIDADE

Este guia, por ser disponibilizado em meio digital, tem abrangência para qualquer leitor em Língua Portuguesa, de reprodutibilidade irrestrita e ampliável, uma vez que novas oportunidades de utilização podem surgir com o avanço tecnológico.

INOVAÇÃO

Trata-se de produto de médio teor inovativo, caracterizado pela organização de conhecimento e definições prévias, com recomendações para facilitar a utilização do ChatGPT como ferramenta educacional para o ensino do raciocínio clínico.

COMPLEXIDADE

Trata-se de produto de baixa complexidade, que pode ser disponibilizado *online* gratuitamente.