RBEM
Revista Brasileira de Educação Médica

ARTIGO ORIGINAL

DOI: https://doi.org/10.1590/1981-5271v49.4-2025-0026

Estudantes de medicina *versus chatbots* na resolução de um teste médico: um estudo comparativo

Medical students versus chatbots in solving a medical test: a comparative study

Douglas Augusto Pasqual La Maison[1] douglaslamaison@gmail.com
Amanda Carolina Fonseca da Silva[1] amandacfsr1@gmail.com
Bruno Ferreira Glir[1] brunofglir@gmail.com
Ari Ojeda Ocampo Moré[1] darimore@hotmail.com

RESUMO

Introdução: A educação médica é avaliada globalmente por Exames Nacionais de Licenciamento Médico e Testes de Progresso. Recentemente, a inteligência artificial generativa (*generative artificial intelligence* – GAI), incluindo *Large Language Models* (LLM) e *chatbots*, mostrou potencial como ferramenta de suporte na educação médica, ao passo que estudos indicaram que seu desempenho é comparável ao de estudantes de Medicina em exames.

Objetivo: Este estudo teve como objetivo avaliar a efetividade de diferentes *chatbots* de LLM na resolução de questões do Teste de Progresso Napisul-II, comparando seu desempenho com o de estudantes do último semestre de Medicina, com vista à aplicabilidade pedagógica de LLM no ensino médico.

Método: Foram utilizados dados do Teste de Progresso Napisul-II de 2023, aplicados em estudantes de Medicina do Sul do Brasil. Quatro *chatbots* (ChatGPT 3.5, ChatGPT 4.0, Bing AI e Bard) responderam três vezes ao teste. As respostas foram comparadas com as dos estudantes do último semestre do curso, de acordo com as áreas médicas inerentes ao TP e as categorias de raciocínio definidas pelos autores. A análise dos dados foi descritiva, utilizando taxas de acerto para comparar os grupos.

Resultado: Os *chatbots* tiveram uma média de 80,81% de acertos em 116 questões, enquanto os estudantes de Medicina obtiveram uma média de 62%. O Bing AI teve a melhor *performance*, com 87,35% de acertos, seguido pelo ChatGPT 4.0 com 85,86%. Em áreas específicas, o ChatGPT 4.0 destacou-se em clínica cirúrgica (88,33%) e clínica médica (86,67%), enquanto o Bing AI teve o melhor desempenho entre os *chatbots* em ciências básicas (94,44%), ginecologia e obstetrícia (94,74%), saúde coletiva (83,33%) e pediatria (82,46%). O Bard, apesar de apresentar a maior amplitude de variação entre as áreas, e o ChatGPT 3.5, inferior aos outros *chatbots*, também superaram os estudantes em todas as categorias.

Conclusão: Todos os *chatbots* estudados tiveram um desempenho superior ao dos discentes de Medicina do 12º semestre do bloco Napisul-II no Teste de Progresso. O Bing destacou-se com o melhor desempenho entre os chatbots, seguido pelo ChatGPT 4.0. Isso indica que os *chatbots* têm potencial de ser uma ferramenta auxiliar na educação médica, embora sejam necessários mais estudos para avaliar as possibilidades dela nesse contexto.

Palavras-chave: Educação Médica; Inteligência Artificial; Desempenho Acadêmico.

ABSTRACT

Introduction: *Medical education is assessed globally by National Medical Licensing Examinations and Progress Tests. Recently, generative artificial intelligence (GAI), including Large Language Models (LLMs) and chatbots, have shown potential as a support tool in medical education, while studies have indicated that their performance is comparable to that of medical students in exams.*

Objective: *To evaluate the effectiveness of different LLM chatbots in solving questions from the NAPISUL-II Progress Test, comparing their performance with that of students in the last semester of medical school, with a view to the pedagogical applicability of LLMs in medical education.*

Method: *Data from the 2023 NAPISUL-II Progress Test, administered to medical students from southern Brazil were used. Four chatbots (ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard) answered the test three times. The responses were compared with those of students in the last semester of the course, according to the medical areas inherent to the PT and reasoning categories defined by the authors. Data analysis was descriptive, using accuracy rates to compare the groups.*

Result: *Chatbots had an average of 80.81% accuracy in 116 questions, while medical students obtained an average of 62%. Bing AI had the best performance, with 87.35% of accuracy, followed by ChatGPT 4.0 with 85.86%. In specific areas, ChatGPT 4.0 stood out in surgical clinic (88.33%) and internal medicine (86.67%), while Bing AI had the best performance among chatbots in basic sciences (94.44%), gynecology and obstetrics (94.74%), public health (83.33%) and pediatrics (82.46%). Bard, despite presenting the greatest range of variation between areas, and ChatGPT 3.5, inferior to the other chatbots, also outperformed students in all categories.*

Conclusion: *All chatbots assessed outperformed medical students in the twelfth semester of the NAPISUL-II block in the Progress Test. Bing stood out with the best performance among the chatbots, followed by Chat GPT-4. This indicates that chatbots have the potential to be an auxiliary tool in medical education, although more studies are needed to evaluate the possibilities of this tool in this context.*

Keywords: *Medical Education; Artificial Intelligence; Academic Performance.*

[1] Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, Brasil.

Editora-chefe: Rosiane Viana Zuza Diniz. | Editor associado: Jorge Guedes.

Recebido em 03/02/25; Aceito em 20/06/25. | Avaliado pelo processo de double blind review.

INTRODUÇÃO

O ensino médico é avaliado por meio de Exames Nacionais de Licenciamento Médico em uma grande gama de países, como Estados Unidos, Alemanha e China. Esses exames buscam avaliar se a *performance* de estudantes de Medicina em competências médicas é suficiente para uma prática de excelência. Para além de uma ferramenta de avaliação de competências, os exames acabam servindo inclusive para guiar a evolução da educação médica, uma vez que seus resultados podem representar a qualidade das universidades e orientá-las para a melhoria de seu ensino. Entretanto, tais exames têm como objetivo principal o licenciamento, e não a avaliação do ensino médico[1-3]. Para tal propósito, desenvolveu-se em 1970, na University of Missouri-Kansas City School of Medicine, nos Estados Unidos, e na University of Limburg, na Holanda, uma ferramenta denominada Teste de Progresso (TP). O teste é constituído por questões de múltipla escolha que avaliam aspectos cognitivos, sendo o mesmo teste aplicado, ao mesmo tempo, para estudantes de todos os níveis do curso de Medicina. O processo de avaliação tem como objetivo compreender a correspondência entre o desempenho dos estudantes e o ano de estudo em que estão matriculados, além de analisar o desenvolvimento do aprendizado de um mesmo aluno ao longo do curso. Dessa maneira, os métodos didáticos são aperfeiçoados a partir da identificação de falhas em determinadas áreas de ensino, visualizadas nas curvas de desempenho acadêmico obtidas por meio dos resultados do TP[4-9].

No Brasil, a aplicação de TP é mais recente e suas questões são baseadas nas Diretrizes Curriculares Nacionais para escolas médicas. Relacionados à Associação Brasileira de Educação Médica (Abem), existem 16 núcleos interinstitucionais que reúnem uma certa quantidade de escolas médicas para aplicação do teste, possibilitando análises mais robustas e comparações. No Sul do país, há o Núcleo de Apoio Pedagógico Interinstitucional Sul II (Napisul-II), composto por seis escolas paranaenses e sete de Santa Catarina. Nesse núcleo, o TP teve sua 13ª edição em outubro de 2023[10].

Recentemente, popularizou-se o uso de *generative artificial intelligence* (GAI), a qual se mostra promissora na área da educação[11-15], e há na literatura pesquisas que envolvem o uso da GAI para resolução de Exames Nacionais de Licenciamento Médico[16,17]. A GAI diz respeito a uma tecnologia capaz de produzir conteúdos com base em bancos de dados prévios[18]. Dentro do universo das IA generativas, encontra-se a família dos *Generative Pre-trained Transformers* (GPT), que são *Large Language Models* (LLM)[19]. Os LLM, por sua vez, são algoritmos de inteligência artificial (IA) projetados para determinar a probabilidade de determinadas sequências de palavras, levando em consideração o contexto das palavras que as antecedem[20]. O aperfeiçoamento desses algoritmos foi revolucionário no campo de *Natural Language Processing* (NLP), um tipo de aprendizado de máquina que simula a linguagem humana[19].

Considerando que a linguagem é um fator-chave no campo da medicina, os LLM mostram-se extremamente promissores em aprender a representar o conhecimento médico. Entretanto, é evidente que essas ferramentas denotam gerações de texto que podem não representar a veracidade da ciência médica ou de sua ética[21]. Mais recentemente, observou-se a popularização dos *chatbots* (ChatGPT, Bing, Bard) – interfaces para sistemas de IA baseados em modelos de linguagem grandes e generativos pré-treinados (GPT), que produzem textos semelhantes aos dos humanos em resposta às mensagens enviadas pelos usuários. Isso possibilitou que tais ferramentas, antes utilizadas apenas por profissionais relacionados às IA, sejam utilizadas pela população leiga, o que coloca sob apreciação a possibilidade e os perigos de seu uso no âmbito da educação médica[19,22,23].

Considerando tais possibilidades, foram desenvolvidas inúmeras pesquisas explorando as capacidades dessas ferramentas na resolução de exames de licenciamento médico, testes específicos de especialidades médicas ou de outras áreas da saúde e TP[16,17,24-26]. Obtiveram-se resultados complementares entre si, uma vez que o ChatGPT 4.0 teve no *United States Medical Licensing Examination* (USMLE) uma *performance* similar à de um estudante de terceiro ano de Medicina[17]. Além disso, o ChatGPT 4.0 se comportou como um residente do primeiro ano de cirurgia plástica[25] e foi aprovado no *Japanese Medical Licensing Examination*[16], o que demonstra seu potencial como instrumento de suporte interativo ao ensino médico.

Evidencia-se uma lacuna na literatura científica no que diz respeito à comparação de desempenho entre múltiplos *chatbots* de LLM e estudantes brasileiros de Medicina, uma vez que já existem estudos semelhantes envolvendo discentes de outras nacionalidades ou limitados a apenas um *chatbot*[27-30]. Apesar de pesquisas internacionais fornecerem *insights* sobre o potencial dos *chatbots* de LLM na educação médica, a variabilidade cultural, linguística e principalmente curricular entre diferentes países sugere que os resultados e as implicações dessas tecnologias podem ser significativamente distintos no Brasil.

Nesse contexto, o objetivo deste trabalho é preencher essa lacuna ao avaliar a efetividade de diferentes *chatbots* de LLM resolverem questões do TP Napisul-II. Isso envolve uma análise comparativa entre as respostas fornecidas por essas ferramentas de IA e as de estudantes do último semestre de Medicina, a fim de comparar o nível de precisão nas respostas. Tal investigação não apenas contribuirá para o entendimento da aplicabilidade dos *chatbots* de LLM na educação médica

brasileira, mas também fornecerá *insights* valiosos sobre as potenciais limitações e vantagens dessas tecnologias emergentes no aprimoramento da qualidade do ensino e da avaliação médica.

MÉTODO

Contexto

O TP é um modelo de prova objetiva de conhecimentos específicos de determinado curso de nível superior, sendo uma estratégia de avaliação individual bastante consolidada na atualidade para evidenciar o processo de aprendizagem do estudante ao longo de sua formação. No Brasil, a Abem é responsável pela elaboração, expansão e aplicação do TP no contexto do curso de graduação em Medicina.

Coleta de dados

Este estudo utilizou como banco de questões o TP elaborado pelo Napisul-II, um dos grupos regionais da Abem. Adotou-se a edição da prova realizada em 18 de outubro de 2023. O documento-base, do qual se transcreveram as questões, nos foi enviado pela coordenação do curso de graduação em Medicina de uma das instituições que constituem o bloco Napisul-II. Um pesquisador transcreveu as questões para um documento editável no Google Docs e, com a ajuda de outros dois pesquisadores, revisou possíveis erros.

Para a realização deste estudo, foi obtida a devida autorização para o uso de dados dos resultados do TP de 2023 para fins de pesquisa. Primeiramente, realizou-se um contato pessoal, e depois se solicitaram por *e-mail* os dados a um dos membros da coordenação do Napisul-II. Em 7 de dezembro de 2023 obteve-se a resposta do *e-mail* com a autorização do uso dos dados. É importante destacar que todos os dados utilizados são anônimos e desprovidos de identificação individual de estudantes ou núcleos específicos, sendo empregados exclusivamente para fins acadêmicos e de pesquisa. Adicionalmente, ressalta-se que esses dados gerais do Napisul-II utilizados nesta pesquisa, além de terem acesso público, são amplamente divulgados no contexto universitário.

Desenho de estudo

De início, foram definidos quatro *chatbots* para análise: ChatGPT 3.5, ChatGPT 4.0, Bing AI e Bard (depois o Bing AI e Bard passaram a se chamar Copilot e Gemini, respectivamente)[31-33] por serem os mais relevantes no momento da coleta[34]. Os *chatbots* foram acessados através de suas interfaces públicas na *web*. Realizaram-se os testes de todos os *chatbots* entre os dias 14 de novembro e 2 de dezembro de 2023. Todas as séries de perguntas foram precedidas pelo seguinte comando-padrão: "A partir de agora, serão enviadas questões de múltipla escolha, e você deve selecionar apenas uma alternativa correta. Seja pontual, marque a alternativa sem explicar o motivo", de forma a seguir o modelo de engenharia de *prompt* para questões médicas *direct one-shot*[35]. Respostas que indicaram mais de uma alternativa como correta foram consideradas como erradas. Dado o comando inicial, as questões foram enviadas de forma individual, sendo separadas em grupos de 20 questões, delimitadas por áreas de conhecimento médico predefinidas pelo TP, com o objetivo de facilitar a organização e evitar que o *chatbot* negligenciasse o comando original por viés de memória[20,36]. Ao término de cada bloco, um novo *chat* era aberto, e enviava-se novamente o comando inicial. Todas as questões do TP foram encaminhadas, com exceção de questões anuladas (duas), com imagens (uma) ou tabelas (uma), dado que, no momento da coleta, existiam diferentes limitações na capacidade de processamento e interpretação visual dos *chatbots*[37]. Após a conclusão dessa etapa, todas questões e suas respectivas respostas foram registradas em um documento na plataforma Google Docs, por meio de capturas de tela e *links* que remetem ao bloco de questões respondidas da área de conhecimento médico apresentada. Todo esse processo foi realizado três vezes, assim como feito em um estudo prévio semelhante[38], para que a confiabilidade dos dados obtidos fosse maior. Essa abordagem de mútiplos ensaios para cada questão é fundamental ao testar LLM, dada a variabilidade inerente em suas respostas, garantindo uma avaliação mais robusta e representativa de seu desempenho potencial. Após isso, criou-se uma tabela por meio do Google Sheets, na qual se evidenciaram as respostas dadas pelos *chatbots* ao lado do gabarito oficial do TP, para posterior avaliação de assertividade.

Análise de dados

Para a análise dos dados, utilizou-se a abordagem de epidemiologia descritiva, adequada para descrever e comparar as taxas de acerto entre diferentes grupos (estudantes e *chatbots*) em diversos contextos. Essa abordagem permite uma avaliação detalhada da distribuição de acertos, facilitando a visualização de padrões e tendências, como as diferenças de desempenho entre os estudantes da 12ª fase do Napisul-II e os *chatbots*. A abordagem de epidemiologia analítica não foi utilizada porque o objetivo principal era descrever e comparar as taxas de acerto entre grupos, sem investigar associações causais entre variáveis. Além disso, não foi o intuito desta pesquisa a demonstração de diferenças estatisticamente significantes, pois o estudo focou a apresentação e interpretação de padrões e distribuições dos dados.

As respostas corretas de cada *chatbot* foram contabilizadas, sendo dividido o número de acertos pelo número de possíveis acertos, assim obtendo a taxa de acerto de

cada *chatbot*. Já que optamos pela resolução do TP por completo três vezes, essa contagem se baseia na média obtida a partir do total de acertos nas três tentativas, dividido pelo número total de chances (348). Já no caso das informações acerca dos estudantes de Medicina do bloco Napisul-II, o banco de dados foi o resultado oficial do TP. Com isso, quando se comparou a porcentagem média de acertos entre os estudantes de Medicina do Napisul-II em geral, os estudantes de Medicina do bloco Napisul-II que estavam na 12ª fase do curso e a porcentagem média de acerto de cada *chatbot*, foi possível ter um panorama geral do desempenho desses três grupos na resolução do TP. Após isso, um enfoque maior foi dado à comparação específica entre os estudantes da 12ª fase do Napisul-II e os *chatbots*, também por meio das médias de acerto, mas com distinção entre os seis blocos definidos pela Abem, isto é: ciências básicas; saúde coletiva; ginecologia e obstetrícia; clínica médica; clínica cirúrgica; pediatria. Posteriormente, com o intuito de ter uma noção maior da forma como estão distribuídos os *chatbots* e os estudantes de Medicina da 12ª fase, indicaram-se os intervalos em que a média de cada *chatbot* ficou. Além disso, fez-se um detalhamento das assertividades por meio de uma tabela, em que foram dispostas as porcentagens de acerto dos estudantes e de cada *chatbot* em relação às áreas do próprio TP: ciências básicas; ginecologia e obstetrícia; saúde coletiva; clínica cirúrgica; clínica médica; pediatria. Como forma de testar a confiabilidade das informações, montou-se um gráfico que facilita a visualização da constância dos *chatbots* ao elucidar a quantidade de vezes que cada *chatbot* acertou zero, uma, duas ou três vezes. Por fim, dois pesquisadores elencaram as questões de acordo com habilidades específicas da prática médica. Definiram-se os grupos de forma a contemplar mais de dez questões cada, garantindo ao menos 30 respostas por *chatbot*, já que todas as questões foram respondidas três vezes por um mesmo *chatbot*. Com isso, definiram-se as seguintes categorias: conduta diagnóstica; conduta terapêutica; conduta terapêutica medicamentosa; conduta terapêutica de procedimento; diagnóstico. O objetivo de cada uma das categorias foi, respectivamente: tomada de decisões diagnósticas, que dependem de diferentes áreas do conhecimento por relacionarem diferentes aspectos do saber médico; tomada de decisões terapêuticas complexas, que envolvem a possibilidade tanto do uso de fármacos quanto da realização de manobras, procedimentos ou mesmo da decisão pela ausência de intervenções; selecionar fármacos adequados à situação, conhecimento também multifatorial, pois depende não somente das condições do paciente, mas também de saberes relacionados à farmacologia; optar corretamente por situações que exijam manejo físico, o que inclui manobras e técnicas cirúrgicas; analisar todos os sinais e sintomas que o paciente apresenta, o contexto em que se encontra, e todas as demais informações apresentadas para relacionar a uma comorbidade. Com isso, uma segunda tabela foi feita, para que cada tipo de conhecimento tivesse sua taxa de acerto por *chatbot* em relação aos estudantes da 12ª fase do Napisul-II evidenciada separadamente.

Considerações éticas

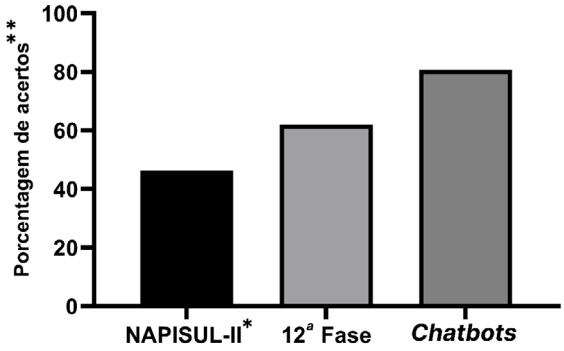
O modelo de pesquisa empregado no presente estudo dispensa avaliação prévia pelo Comitê de Ética em Pesquisa (CEP), pois os dados referentes ao desempenho acadêmico do bloco Napisul-II na realização do TP são anônimos e estão acessíveis ao público geral[39].

RESULTADOS

Os *chatbots* tiveram uma pontuação média de 93,73 (80,81%) acertos em 116 questões, ultrapassando os estudantes de Medicina da 12ª fase do bloco Napisul-II, que obtiveram uma média de 74,4 (62%) acertos em 120 questões, seguidos pelos estudantes de Medicina em geral do bloco Napisul-II, com 55,58 (46,31%) acertos em 120 questões.

Em uma análise mais aprofundada, foi perceptível que o Bing demonstrou uma *performance* superior a todos outros, com uma pontuação média de 101,33 de 116 (87,35%), ultrapassando o ChatGPT 4.0 que alcançou 99,6 de 116 (85,86%). Em sequência, estão, respectivamente: Bard com uma média de 89 de 116 (76,72%) e ChatGPT 3.5 com uma média de 85 de 116 (73,27%). Essas informações foram comparadas à frequência de estudantes em cada faixa de acerto com o uso de figuras geométricas representando a faixa de acerto da média de cada *chatbot*.

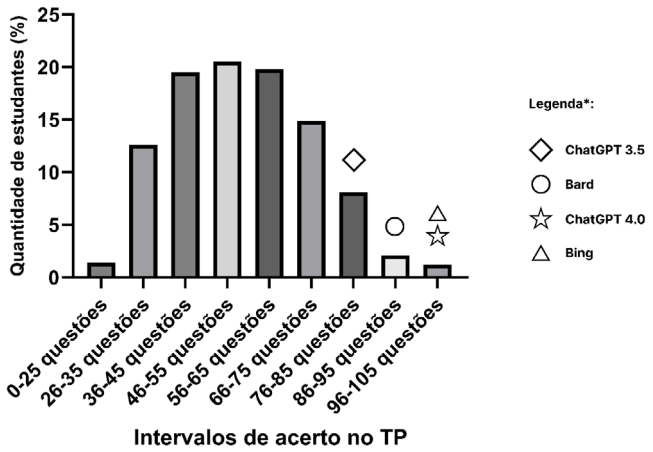
Gráfico 1. Porcentagem de acertos no Teste de Progresso de todos os alunos do Napisul-II, dos estudantes de 12ª fase do Napisul-II e dos *chatbots*.



* A média de acertos do Napisul-II corresponde a todos os estudantes do bloco Napisul-II.
** As médias do Napisul-II e da 12ª fase foram coletadas dos resultados oficiais do TP, em que constavam 120 questões; já a média dos *chatbots* desconsiderou duas questões anuladas e duas questões com tabela ou imagem, sendo analisada quanto às 116 questões restantes.
Fonte: Elaborado pelos autores.

 DOI: https://doi.org/10.1590/1981-5271v49.4-2025-0026

Gráfico 2. Distribuição dos estudantes do Napisul-II e dos *chatbots* por intervalo de acerto no Teste de Progresso.



* Os símbolos representam em qual intervalo de acertos esteve o desempenho médio de *cada chatbot*.
Fonte: Elaborado pelos autores.

Na divisão por áreas, isto é, ciências básicas, ginecologia e obstetrícia, saúde coletiva, clínica cirúrgica, clínica médica e pediatria, os estudantes tiveram apenas uma área com desvio > 5% em relação à média geral: ginecologia e obstetrícia, com 67,2% de acerto, sendo a maior taxa de acerto alcançada por eles. O ChatGPT 3.5, embora tenha sido o *chatbot* com a pior média de acertos, manteve uma média maior que os estudantes em todas as áreas. No caso do ChatGPT 4.10, houve destaques em clínica cirúrgica (88,33%) e clínica médica (86,67%), áreas nas quais pontuou mais que qualquer outro *chatbot*. O Bing, o *chatbot* com maior média geral, alcançou a maior média em ciências básicas (94,44%), ginecologia e obstetrícia (94,74%), saúde coletiva (83,33%) e pediatria (82,46%). O Bard foi o *chatbot* com maior inconstância de médias, alcançando 92,59% em ciências básicas e 66,67% em pediatria. A Tabela 1 apresenta essas informações.

Tabela 1. Comparativo de percentuais de acerto entre os estudantes da 12ª fase do curso de graduação em Medicina das faculdades que compõem o Napisul-II e os *chatbots* no Teste de Progresso quanto às divisões por área inerente

Área do TP	12ª fase	ChatGPT 3.5	ChatGPT 4.0	Bing	Bard
Ciências básicas	60,50%	81,48%	88,89%	94,44%	92,59%
Ginecologia e obstetrícia	67,20%	75,44%	92,98%	94,74%	71,93%
Saúde coletiva	59,40%	61,67%	80,00%	83,33%	75,00%
Clínica cirúrgica	62,10%	75,00%	88,33%	85,00%	80,00%
Clínica médica	62,50%	76,67%	86,67%	85,00%	73,33%
Pediatria	60,40%	66,67%	78,95%	82,46%	66,67%
Total	62,00%	73,27%	85,91%	87,36%	76,72%

Fonte: Elaborada pelos autores.

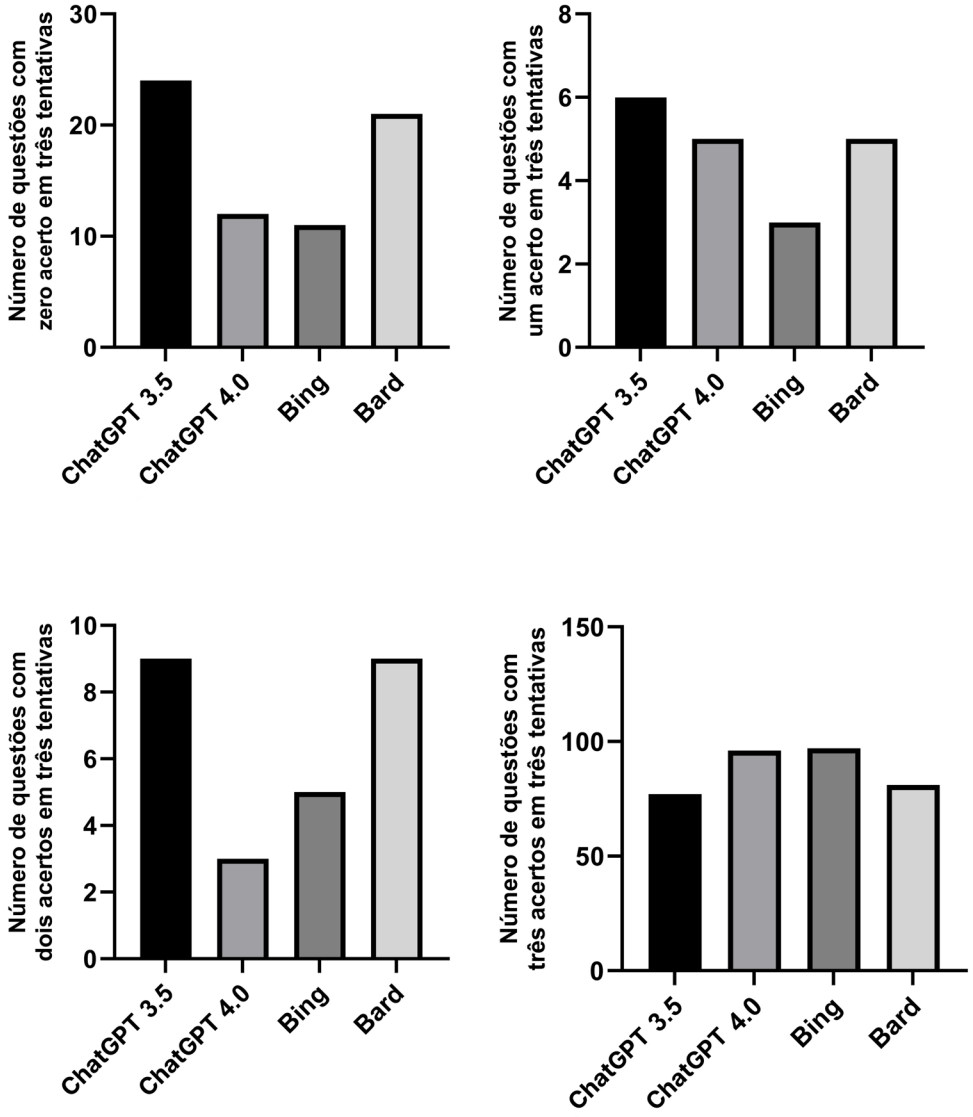
Tabela 2. Comparativo de percentuais de acerto entre os estudantes da 12ª fase do curso de graduação em Medicina das faculdades que compõem o Napisul-II e os *chatbots* quanto aos principais tipos de conhecimentos médicos abordados no Teste de Progresso.

Tipo de conhecimento	Número de questões*	12ª fase	ChatGPT 3.5	ChatGPT 4.0	Bing	Bard
Conduta diagnóstica	16	54,53%	72,92%	71,88%	66,67%	60%
Conduta terapêutica	36	56,81%	70,37%	85,19%	83,33%	67,59%
Conduta terapêutica (medicamentosa)	14	51,91%	73,81%	92,86%	100%	61,90%
Conduta terapêutica (procedimento)	11	61,33%	69,70%	84,85%	75,76%	63,64%
Diagnóstico	24	60,66%	78,26%	88,41%	91,30%	76,81%

* O número de tentativas em análise é três vezes o número de questões, pois cada *chatbot* realizou o teste três vezes.
Fonte: Elaborada pelos autores.

 DOI: https://doi.org/10.1590/1981-5271v49.4-2025-0026

Gráfico 3. Número de questões com zero, um, dois e três acertos em três tentativas por *chatbot* no Teste de Progresso Napisul-II.



Fonte: Elaborado pelos autores.

DISCUSSÃO

Este estudo avança significativamente em relação à literatura previamente publicada sobre o uso de LLM no contexto do TP brasileiro. Enquanto o estudo de Rodrigues Alessi et al.[28] avaliou exclusivamente o desempenho do ChatGPT 3.5 em relação ao TP, o presente trabalho expande o escopo ao incluir múltiplos modelos de IA (ChatGPT 3.5, ChatGPT 4.0, Bing AI e Bard), introduzindo uma abordagem comparativa mais robusta. Essa ampliação permite não apenas confirmar os achados anteriores, como também identificar variações de desempenho entre diferentes ferramentas de IA generativa, contextualizando melhor suas possíveis aplicações pedagógicas.

Os resultados deste estudo destacam a notável eficácia dos *chatbots* de LLM na resolução de questões médicas, com o Bing AI prevalecendo sobre os demais com uma taxa de acerto de 87,35%, seguido pelo ChatGPT 4.0, com 85,86%. Esse desempenho foi superior em comparação aos estudantes do 12º período do Napisul-II, que apresentaram uma taxa de acerto de 62%. Esses resultados foram ao encontro de estudos anteriores, que demonstraram superioridade do Bing AI e ChatGPT 4.0 sobre estudantes da área da saúde de diferentes níveis e outros *chatbots*, até mesmo aqueles não explorados neste estudo, tais como o Claude Instant e Claude. Em comparação com os resultados obtidos por Rodrigues Alessi et al.[28], o presente estudo observa um aumento geral nas taxas de acerto dos modelos de linguagem. Enquanto o ChatGPT 3.5 apresentou média de 67,2% a 69,7% de acerto nos anos analisados anteriormente, no presente estudo obteve 73,27% em 2023. O ChatGPT 4.0 e o Bing AI, não incluídos na análise anterior, superaram amplamente esse desempenho, com médias de 85,86% e 87,35%, respectivamente. Esses dados

sugerem não apenas evolução tecnológica dos modelos, mas também maior estabilidade de resposta e capacidade de generalização clínica, principalmente quando avaliados em múltiplas tentativas. Além disso, ao adotar uma metodologia que contempla a repetição tripla das questões por cada modelo e a análise por áreas de conhecimento e tipos de raciocínio clínico, este trabalho oferece uma granularidade analítica superior àquela do estudo anterior. Essa abordagem permite inferências mais precisas sobre o tipo de competência médica que pode ser reforçada com o uso de *chatbots* – por exemplo, a superioridade dos modelos em questões de conduta terapêutica medicamentosa e diagnóstico clínico, áreas fundamentais para a prática médica. Assim, este estudo contribui com uma base mais sólida para o desenvolvimento de estratégias educacionais baseadas em IA[28]. No estudo de Morreel et al.[40], o Bing AI e o ChatGPT 4.0 obtiveram, no Exame de Licenciamento Médico de Múltipla Escolha da Antwerp University, 76% de acerto, enquanto os outros *chatbots* avaliados obtiveram 62%-67% e os estudantes 61%, ao passo que Torres-Zegarra et al.[38] encontraram, no Exame de Licenciamento Médico de Múltipla Escolha do Peru, um resultado de 82,2% por parte do Bing, 86,7% pelo ChatGPT 4.0, com o Bard e ChatGPT 3.5 apresentando menor desempenho (ambos 68,8%), mas ainda superior ao resultado obtido pelos estudantes (55%).

Além disso, a análise por áreas específicas revelou que os *chatbots* mantiveram uma *performance* superior aos estudantes em várias disciplinas médicas. Por exemplo, o ChatGPT 4.0 obteve, em ciências básicas, ginecologia e obstetrícia, saúde coletiva, clínica cirúrgica, clínica médica e pediatria, os desempenhos de 88,89%, 92,98%, 80%, 88,33%, 86,67% e 78,95%, respectivamente, ao passo que os estudantes da 12ª fase obtiveram 60,50%, 67,20%, 59,40%, 62,10%, 62,50% e 60,40%. Esses resultados indicam que os *chatbots* não apenas superam os estudantes em média geral, mas também demonstram superioridade em áreas específicas do conhecimento médico. Ademais, o Bing demonstrou superioridade em relação aos outros *chatbots* em todas as áreas médicas avaliadas no TP, excetuando-se clínica médica e clínica cirúrgica, nas quais houve superioridade do ChatGPT4.0. Tal dado divergiu em relação a Torres-Zegarra et al.[38], em que o Bing ficou atrás do ChatGPT 4.0 também nas áreas de pediatria e saúde coletiva, porém apresentando mesma quantidade de acertos em clínica médica.

Outrossim, vale ressaltar que a demonstração de altas taxas de acerto por parte dos *chatbots*, independentemente da área médica ou do tipo de conhecimento avaliado, é presente em estudos realizados em diversos idiomas, países e formas de cobrança de conteúdo, como se pode observar em estudos prévios de Friederichs et al.[29], Meo et al.[41], Tong et al.[42], Aljindan et al.[43], Ebrahimian et al.[44] e Lai et al.[45].

Uma das principais limitações deste estudo reside no fato de que a amostra utilizada, composta por estudantes de medicina do 12º período de um único núcleo interinstitucional (Napisul-II), pode não ser representativa de todos os alunos de Medicina no Brasil ou em outros contextos educacionais. A variabilidade nos currículos médicos e nas práticas pedagógicas pode influenciar significativamente o desempenho dos estudantes em testes padronizados, sugerindo a necessidade de amostras mais amplas e diversificadas para validar os achados.

Uma questão específica relacionada ao desempenho dos *chatbots* é o possível viés de memória entre diferentes tentativas da mesma questão. Embora as questões tenham sido apresentadas em grupos de 20 e precedidas por um comando-padrão para evitar viés de memória, a capacidade dos *chatbots* de armazenar informações de sessões anteriores pode ter influenciado seus resultados. *Chatbots* mais avançados, como o ChatGPT 4.0, podem ter um mecanismo de memória mais sofisticado que lhes permita lembrar respostas anteriores, potencialmente influenciando a precisão das respostas em tentativas subsequentes.

Além disso, outro viés significativo neste estudo é a diferença na capacidade de acesso à internet entre os *chatbots* testados. Alguns *chatbots*, como o Bing AI, têm a capacidade de realizar pesquisas em tempo real na internet para fornecer respostas atualizadas, enquanto outros, como o ChatGPT 3.5 e o ChatGPT 4.0, dependem exclusivamente de seus dados de treinamento preexistentes sem acesso à internet. Essa discrepância pode ter dado ao Bing AI uma vantagem injusta, permitindo-lhe acessar informações mais recentes e potencialmente mais precisas, enquanto os outros modelos poderiam estar limitados por dados desatualizados ou incompletos.

Outra limitação importante a ser considerada neste estudo é que o TP, utilizado como ferramenta de avaliação, é aplicado em estudantes de Medicina sem oferecer benefícios diretos, o que pode influenciar negativamente o desempenho dos alunos. Ao contrário de exames de licenciamento médico, em que a obtenção da licença profissional depende do resultado e, portanto, motiva os estudantes a se empenhar ao máximo, o TP pode não engajar os participantes da mesma forma, resultando em um desempenho que não reflete suas reais capacidades e conhecimentos. Isso pode levar a uma subestimação da competência dos alunos, afetando a comparabilidade dos resultados com estudos que envolvem exames de licenciamento, em que a motivação é significativamente maior e o desempenho tende a ser mais representativo das habilidades dos candidatos.

Contudo, este estudo apresenta uma base sólida para futuras investigações e para o desenvolvimento de diretrizes para a integração de IA no ensino médico, uma vez que faz análises comparativas não apenas no que diz respeito à totalidade de acertos, mas também no que concerne ao desempenho relacionado às diferentes áreas médicas e a tipos de raciocínio a serem desenvolvidos, o que pode ser fundamental para o desenvolvimento de futuras ferramentas direcionadas para habilidades específicas.

Afinal, os *chatbots* baseados em LLM têm um potencial significativo para auxiliar no ensino médico. Kung et al.[20] demonstraram que o ChatGPT forneceu explicações altamente coerentes e *insights* valiosos que facilitam o processo de aprendizagem. As respostas apresentaram alta concordância interna e modelaram um raciocínio dedutivo, ao introduzirem conceitos novos e não óbvios aos estudantes, o que enriquece a compreensão e o raciocínio clínico. De forma semelhante, Cascella et al.[46] evidenciaram a capacidade do ChatGPT em auxiliar na elaboração de notas médicas e na comunicação de informações complexas de maneira compreensível. O *chatbot* categorizou corretamente parâmetros clínicos apresentados de forma desordenada ou abreviada, forneceu sugestões relevantes para tratamentos adicionais e adaptou a linguagem conforme o público, seja entre profissionais de saúde ou pacientes. Essas evidências, somadas às obtidas em nosso estudo, sugerem que a integração de *chatbots* no ensino médico pode complementar os métodos tradicionais, ao oferecer suporte personalizado, promover o desenvolvimento de habilidades clínicas e facilitar a compreensão de materiais complexos, contribuindo para um processo de aprendizagem mais autônomo e eficaz.

CONCLUSÃO

Os resultados deste estudo evidenciam que os *chatbots* avaliados, especialmente o Bing AI e o ChatGPT 4.0, apresentam uma capacidade superior na resolução de questões de avaliação médica em comparação aos estudantes do último semestre de Medicina. Essa constatação não apenas reforça o papel potencial das ferramentas de IA como complementos valiosos ao ensino tradicional, mas também abre um horizonte promissor para a integração dessas tecnologias em práticas educacionais que visem ao aprimoramento do aprendizado e da avaliação médica.

A análise segmentada das áreas de conhecimento destacou que os *chatbots*, além de superarem os estudantes em termos gerais, demonstraram um desempenho consistente em áreas críticas, como ciências básicas, ginecologia e obstetrícia, e clínica cirúrgica. Esses dados sugerem que a aplicação dos *chatbots* pode ser particularmente útil em reforçar o ensino em disciplinas em que os estudantes enfrentam maiores desafios, oferecendo, assim, um suporte significativo ao processo de formação médica. Ademais, a superioridade observada em áreas de conduta terapêutica e diagnóstica aponta para a viabilidade de integrar essas ferramentas em cenários de simulação clínica, ampliando o leque de recursos disponíveis para a educação médica.

Entretanto, é essencial reconhecer as limitações permeiam este estudo, como o viés de memória dos *chatbots* e a motivação reduzida dos estudantes no TP, o que ressalta a necessidade de uma implementação cuidadosa e criteriosa dessas tecnologias. Para que os benefícios identificados sejam plenamente aproveitados, futuras pesquisas devem se concentrar em estratégias para mitigar esses vieses e em adaptações dos *chatbots* às especificidades curriculares e culturais das diferentes instituições de ensino. Dessa forma, com uma abordagem estratégica e contextualizada, os *chatbots* podem se tornar potenciais aliados na formação de médicos mais bem preparados para os atuais desafios da prática clínica.

CONTRIBUIÇÃO DOS AUTORES

Douglas Augusto Pasqual La Maison participou da idealização colaborativa do tema do artigo, da delimitação da metodologia do trabalho, do envio de grande parte das questões aos *chatbots* e do registro das respectivas respostas, da transcrição das respostas para uma tabela, da construção de gráficos e tabelas presentes no texto, da escrita da metodologia e dos resultados, da revisão de literatura e da formatação do artigo. Amanda Carolina Fonseca da Silva participou da idealização conjunta do tema do artigo, da delimitação da metodologia do trabalho, do desenvolvimento do *prompt*, da solicitação dos usos dos dados do Napisul-II, do envio das questões aos *chatbots* e do registro das respectivas respostas, da escrita da introdução e da discussão, da construção de gráficos, da revisão de literatura e da organização das referências do trabalho. Ari Ojeda Ocampo Moré idealizou e orientou do trabalho, e participou de todas as etapas da pesquisa e da escrita científica. Bruno Ferreira Glir participou da coleta de dados, da revisão de literatura referente à metodologia e da revisão da formatação do artigo.

CONFLITO DE INTERESSES

Declaramos não haver conflito de interesses.

FINANCIAMENTO

Declaramos não haver financiamento.

DECLARAÇÃO DE DISPONIBILIDADE DE DADOS

Os dados de pesquisa estão disponíveis no corpo do documento.
Research data is available in the body of the document.

REFERÊNCIAS

1. Wang X. Experiences, challenges, and prospects of National Medical Licensing Examination in China. BMC Med Educ. 2022;22(1):349.
2. Huber-Lang M, Palmer A, Grab C, Boeckers A, Boeckers TM, Oechsner W. Visions and reality: the idea of competence-oriented assessment for German medical students is not yet realised in licensing examinations. GMS J Med Educ. 2017;34(2):Doc25 [acesso em 13 jan 2024]. Disponível em: http://www.egms.de/en/journals/zma/2017-34/zma001102.shtml.
3. Babla K, Crampton P, Kronfli M. National licensing examinations: what are they good for? Clin Teach. 2020;17(3):323-5.
4. Reberti AG, Monfredini NH, Ferreira Filho OF, Andrade DFD, Pinheiro CEA, Silva JC. Progress Test in medical school: a systematic review of the literature. Rev Bras Educ Med. 2020;44(1):e014.
5. Sakai MH, Ferreira Filho OF, AImeida MJD, Mashima DA, Marchese MDC. Teste de Progresso e avaliação do curso: dez anos de experiência da medicina da Universidade Estadual de Londrina. Rev Bras Educ Med. 2008;32(2):254-63.
6. Bollela VR, Borges MDC, Troncon LEDA. Avaliação somativa de habilidades cognitivas: experiência envolvendo boas práticas para a elaboração de testes de múltipla escolha e a composição de exames. Rev Bras Educ Med. 2018;42(4):74-85.
7. Sakai MH, Ferreira Filho OF, Matsuo T. Avaliação do crescimento cognitivo do estudante de Medicina: aplicação do teste de equalização no Teste de Progresso. Rev Bras Educ Med. 2011;35(4):493-501.
8. Baldim TL, Arcuri MB, Aparecida C. O TESTE DE PROGRESSO SOB A VISÃO DO DISCENTE. Revista da Faculdade de Medicina de Teresópolis [Internet]. 2018 [cited 2024 Aug 12];2(1):41–54. Available from: https://revista.unifeso.edu.br/index.php/faculdadedemedicinadeteresopolis/article/view/608
9. Presta PM, Oliveira A de P, Moreira MLRM, Costa GOF da, Peixoto RAC. Conhecimento em clínica médica: resultados de Teste de Progresso de uma instituição do Nordeste. Revista Interagir. 2024;(126):36-43.
10. Rosa MID, Isoppo CC, Cattaneo HD, Madeira K, Adami F, Ferreira Filho OF. O Teste de Progresso como indicador para melhorias em curso de graduação em Medicina. Rev Bras Educ Med. 2017;41(1):58-68.
11. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. EdArXiv; 2023 [acesso em 1º fev 2024]. Disponível em: https://osf.io/5er8f.
12. Rasul T, Nair S, Kalendra D, Robin M, Santini FO, Ladeira WJ, et al. The role of ChatGPT in higher education: benefits, challenges, and future research directions. JALT. 2023;6(1):41-56 [acesso em 1º fev 2024]. Disponível em: https://journals.sfu.ca/jalt/index.php/jalt/article/view/787.
13. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare. 2023;11(6):887.
14. Rudolph, J., Tan, S., Tan, S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? JALT. 2023;6(1): 342-362 [acesso em 1º fev 2024]. Disponível em: https://journals.sfu.ca/jalt/index.php/jalt/article/view/689.
15. Costa MJM, Santos DWD, Bottentuit Junior JB. Inteligência artificial e metodologias ativas no ensino de medicina: percepções dos discentes de habilidades médicas de um centro universitário. Revista Intersaberes. 2024;19;e24do3003.
16. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. JMIR Med Educ. 2023;9:e48002.
17. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312.
18. Shoja M M, Van de Ridder J, Rajput V (June 24, 2023) The Emerging Role of Generative Artificial Intelligence in Medical Education, Research, and Practice. Cureus 15(6): e40883. doi:10.7759/cureus.40883.
19. Ramos ASM. Inteligência Artificial Generativa baseada em grandes modelos de linguagem - ferramentas de uso na pesquisa acadêmica [Internet]. SciELO Preprints. 2023 [citado 1º de fevereiro de 2024]. Disponível em: https://preprints.scielo.org/index.php/scielo/preprint/view/6105
20. 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. PLOS Digit Health. 2023;2(2):e0000198.
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
22. Lee H. The rise of ChatGPT : exploring its potential in medical education. Anat Sci Educ. 2024;17:926–931. doi: 10.1002/ase.2270.
23. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9.
24. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc. 2023;86(7):653-8.
25. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service examination. Aesthet Surg J. 2023;43(12):NP1085-9.
26. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European exam in core cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023;4(3):279-81.
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;183(9):1028-1030.
28. Alessi MR, Gomes HA, Castro ML de, Okamoto CT. Performance of ChatGPT in solving questions from the Progress Test (Brazilian National Medical Exam): a potential artificial intelligence tool in medical practice. 2024 Jul 19;16(7):e64924. [acesso em 25 maio 2025]. Disponível em: https://www.cureus.com/articles/272161-performance-of-chatgpt-in-solving-questions-from-the-progress-test-brazilian-national-medical-exam-a-potential-artificial-intelligence-tool-in-medical-practice.
29. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Educ Online. 2023;28(1):2220920.
30. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the possible use of AI chatbots in public health education: feasibility study. JMIR Med Educ. 2023;9:e51421.
31. Chat GPT 4.0 [acesso em Novembro de 2023]. Disponível em: https://chatgpt.com.
32. Copilot AI [acesso em Novembro de 2023]. Disponível em: https://copilot.microsoft.com.
33. Gemini AI [acesso em Novembro de 2023]. Disponível em: https://gemini.google.com.
34. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod 2024;46: cjae017.
35. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? arXiv; 2023 [acesso em 3 jul 2024]. Disponível em: http://arxiv.org/abs/2207.08143.
36. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT–3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. J Am Acad Orthop Surg. 2023;31(23):1173-9.
37. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589-597.
38. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30.
39. Resultados Teste de Progresso [acesso em Julho de 2024]. Disponível em: https://medicina.ufsc.br/?page_id=2462.

40. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. PLOS Digit Health. 2024;3(2):e0000349.

41. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare. 2023;11(14):2046.

42. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese National Medical Licensing Examination. Front Med. 2023;10:1237432.

43. Aljindan FK, Al Qurashi AA, Albalawi IAS, et al. ChatGPT conquers the Saudi medical licensing exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus. 2023 Sep;15(9):e45043. doi: 10.7759/cureus.45043. doi. Medline

44. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform. 2023;30(1):e100815.

45. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom Medical Licensing Assessment. Front Med. 2023;10:1240915.

46. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst. 2023;47(1):33.