



Universidade Federal De Santa Catarina

Centro De Ciências Da Saúde

Departamento De Clínica Médica

Curso de Graduação em Medicina

Douglas Augusto Pasqual La Maison

**ESTUDANTES DE MEDICINA VS CHATBOTS NA RESOLUÇÃO DE
UM TESTE MÉDICO: UM ESTUDO COMPARATIVO**

Florianópolis

2025

DOUGLAS AUGUSTO PASQUAL LA MAISON

**ESTUDANTES DE MEDICINA VS CHATBOTS NA
RESOLUÇÃO DE UM TESTE MÉDICO
um estudo comparativo**

**Trabalho apresentado à Universidade
Federal de Santa Catarina, como requisito
para a conclusão do Curso de Graduação em
Medicina.**

**Florianópolis
Universidade Federal de Santa Catarina
2025**

DOUGLAS AUGUSTO PASQUAL LA MAISON

**ESTUDANTES DE MEDICINA VS CHATBOTS NA
RESOLUÇÃO DE UM TESTE MÉDICO
um estudo comparativo**

**Trabalho apresentado à Universidade
Federal de Santa Catarina, como requisito
para a conclusão do Curso de Graduação em
Medicina.**

Presidente do Colegiado: Prof. Dr. Antônio Reis de Sá Júnior

Professor Orientador: Prof. Dr. Ari Ojeda Ocampo Moré

**Florianópolis
Universidade Federal de Santa Catarina
2025**

La Maison, Douglas Augusto Pasqual

Estudantes de Medicina vs Chatbots na Resolução de um
Teste Médico : Um Estudo Comparativo / Douglas Augusto
Pasqual La Maison ; orientador, Ari Ojeda Ocampo Moré,
2025.

35 p.

Trabalho de Conclusão de Curso (graduação) -
Universidade Federal de Santa Catarina, Centro de Ciências
da Saúde, Graduação em Medicina, Florianópolis, 2025.

Inclui referências.

1. Medicina. 2. Educação Médica . 3. Inteligência
Artificial . 4. Desempenho Acadêmico . I. Moré, Ari Ojeda
Ocampo . II. Universidade Federal de Santa Catarina.
Graduação em Medicina. III. Título.

DEDICATÓRIA

Dedico este trabalho aos meus pais, Mário e Vanessa, que são a razão de tudo o que sou e conquisto. Pela coragem, pelo amor incansável e por me ensinarem, com o próprio exemplo, que não há sonho grande demais para quem carrega dignidade, trabalho e afeto no coração. Tudo o que construo é, antes de tudo, para honrar vocês. E para mim, honrá-los é fazer valer a pena cada passo que me permitiram dar.

(Baseado em Êxodo 20:12)

AGRADECIMENTOS

Há caminhos que parecem improváveis até que a gente os percorra. E, olhando hoje para este trabalho, sei que ele não é fruto apenas de livros ou esforço acadêmico. Ele é, sobretudo, fruto de uma história de amor e memórias.

Meus pais, Mário e Vanessa, são as colunas que sustentam tudo que sou. Eles me ensinaram, desde cedo, que caráter e generosidade constroem qualquer futuro. É impossível expressar em palavras o quanto sou grato a vocês. Mário, é um dos maiores exemplos que carrego na vida. Parceiro de todas as horas, ele sempre esteve ao meu lado. Foi com ele que aprendi o que significa ser homem — não no sentido vazio que a sociedade tenta impor, mas no sentido mais verdadeiro da palavra: ser homem é ser íntegro, é ser justo, é proteger quem se ama. Nosso amor pelo Grêmio é uma herança dele, assim como os ensinamentos das conversas e inúmeros momentos que compartilhamos. Ele sempre me ofereceu suporte para que eu pudesse voar mais alto, sem que eu sequer precisasse pedir. Meu pai é, sem dúvida, uma das maiores bênçãos que a vida me deu. À Vanessa, cuja força e ternura moldaram boa parte do homem que sou hoje, minha melhor amiga que, com uma simples conversa, transforma qualquer dia difícil em leveza. Cresci ouvindo suas histórias, suas risadas e apreciando de perto como ela torna qualquer desafio em coragem para seguir. Com ela aprendi a lidar com a vida como ela é, entendendo não só como cuidar de mim, mas principalmente de quem está ao meu redor. Trago comigo também as marcas dos meus avós, que foram muito mais do que avós: foram presença, casa e amor. Seu Mário, com momentos inesquecíveis jogando sinuca e assistindo o que ele chamava de “bang bang”. Dona Glória, com suas muitas histórias de vida e que sempre me acolheu com afeto. Vanda, que foi puro cuidado e zelo, me mimando com amor em cada detalhe. E Paulo, que alimentou minha curiosidade pelo mundo e me encheu de carinho nas pequenas coisas. À minha irmã, Maria Antônia, deixo registrado todo o meu amor. A vida, com seus caminhos e desencontros, por vezes nos impôs a distância, mas a verdade é que o carinho e o orgulho que sinto por você são imensos. Aos meus amigos, que são uma extensão da minha família. Pedro Sebben, o amigo que sempre esteve na torcida pelo meu sucesso, assim como eu estive pelo dele. A vida acabou nos levando, curiosamente, para a mesma universidade, e hoje seguimos caminhando lado a lado, cada um na sua área, mas sempre compartilhando vitórias, dificuldades e sonhos. Ao Gabriel Muller, amigo de longe que a vida me trouxe de forma improvável, e que hoje levo além das telas. Ao Gabriel Marques, que é muito mais que primo, é irmão de vida. Crescemos juntos desde que éramos praticamente bebês. Fomos parceiros inseparáveis na infância, nas madrugadas jogando, rindo e criando memórias que levarei comigo para sempre. À Amanda, minha companheira. Você trouxe sentido e alívio aos últimos anos da minha vida. Foi com você que aprendi, de forma definitiva, que amor também é comunicação, é aprendizado, é caminhar juntos. Sua presença e a de sua família na minha vida foram fundamentais não só nos dias felizes, mas, principalmente, naqueles em que tudo parecia mais difícil. Seu apoio, sua inteligência e sua capacidade de trazer segurança foram combustíveis indispensáveis para que eu chegasse até aqui de forma mais leve. É você quem me faz querer ser melhor todos os dias. Ao professor Ari Moré, meu mais profundo agradecimento. Foi sob sua orientação que não apenas desenvolvi este trabalho, mas também que fui introduzido ao fascinante mundo das inteligências artificiais na medicina. Sua generosidade e sua dedicação foram fundamentais em cada etapa deste processo. Mais do que um orientador acadêmico, você se tornou uma referência profissional e pessoal, alguém que me inspira a seguir buscando conhecimento, inovação e excelência. Aos professores que compõem esta banca, deixo aqui minha admiração e gratidão. Tê-los presentes nesta etapa é, além de uma honra, uma grande alegria. São profissionais que admiro, e cuja trajetória me inspira — tanto na medicina quanto, quem sabe, no caminho do ensino. Por fim, à minha grande família — os La Maison e os Pasqual — que, de formas diferentes, sempre fizeram questão de estar presentes, torcendo, apoiando e celebrando cada passo da minha vida. Obrigado tios e primos. Que este trabalho seja uma pequena forma de retribuir tudo que recebi até aqui.

"A medicina é uma ciência da incerteza e uma arte da probabilidade."

— William Osler

ESTUDANTES DE MEDICINA VS CHATBOTS NA RESOLUÇÃO DE UM TESTE MÉDICO: UM ESTUDO COMPARATIVO

MEDICAL STUDENTS VS CHATBOTS IN SOLVING A MEDICAL TEST: A COMPARATIVE STUDY

RESUMO

INTRODUÇÃO: A educação médica é avaliada globalmente por Exames Nacionais de Licenciamento Médico e Testes de Progresso. Recentemente, a inteligência artificial generativa (GAI), incluindo Large Language Models (LLMs) 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:** 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 LLMs 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 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 aos estudantes de medicina do décimo segundo semestre do bloco NAPISUL-II no Teste de Progresso. O Bing destacou-se com o melhor desempenho entre os chatbots, seguido pelo Chat GPT-4. Isso

indica que os chatbots tem potencial de ser uma ferramenta auxiliar na educação médica, embora sejam necessários mais estudos para avaliar as possibilidades dessa ferramenta 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:** Were used data from the 2023 NAPISUL-II Progress Test, administered to medical students from Southern Brazil. Four chatbots (ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard) responded to 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% accuracy, followed by ChatGPT 4.0 with 85.86%. In specific areas, ChatGPT 4.0 stood out in surgical clinic (88.33%) and medical clinic (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 studied 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. INTRODUÇÃO

O ensino médico é avaliado por meio de Exames Nacionais de Licenciamento Médico em uma grande gama de países, tais quais 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 melhoria de seu ensino. Entretanto, tais exames têm como objetivo principal o licenciamento, e não a avaliação do ensino médico¹⁻³. 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. 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 Teste de Progresso⁴⁻⁹.

No Brasil, a aplicação de Testes de Progresso é mais recente e suas questões são baseadas nas diretrizes curriculares nacionais para escolas médicas. Relacionadas a ABEM (Associação Brasileira de Educação Médica), 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. Neste núcleo, o teste de progresso teve sua 13ª edição em outubro de 2023¹⁰.

Recentemente popularizou-se o uso de Generative Artificial Intelligence (GAI), a qual se mostra promissora na área da educação¹¹⁻¹⁵, havendo na literatura pesquisas envolvendo seu uso 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 prévios¹⁸. Dentro do universo das IA generativas encontra-se a família dos GPTs (Generative Pre-trained Transformers), que são Large Language Models (LLMs)¹⁹. LLMs, 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 antecedem²⁰. O aperfeiçoamento desses algoritmos foi revolucionário no campo de Natural Language Processing (NLP), um tipo de aprendizado de máquina que simula a linguagem humana¹⁹.

Considerando que a linguagem é um fator chave no campo da medicina, as LLMs mostram-se extremamente promissoras 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 ética²¹. 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 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 testes de progresso^{16,17,24–26}. Foram obtidos resultados complementares entre si, uma vez que o ChatGPT 4.0 obteve no United States Medical Licensing Examination (USMLE) uma performance similar a um estudante de terceiro ano de medicina¹⁷, além de ter se comportado como um residente do primeiro ano de cirurgia plástica²⁵ e ter sido aprovado no Japanese Medical Licensing Examination¹⁶, o que demonstra seu potencial como instrumento de suporte interativo ao ensino médico.

Evidencia-se uma lacuna na literatura científica no que diz respeito a comparação de desempenho entre múltiplos Chatbots de LLM e estudantes brasileiros de medicina, uma vez que já existem estudos semelhantes envolvendo estudantes 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 implicações dessas tecnologias podem ser significativamente diferentes no Brasil.

Nesse contexto, o objetivo deste trabalho é preencher essa lacuna ao avaliar a efetividade de diferentes Chatbots de LLM em resolver questões do teste de progresso 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 os potenciais limitações e vantagens dessas tecnologias emergentes no aprimoramento da qualidade do ensino e da avaliação médica.

2. METODOLOGIA

2.1. Contexto

O teste de progresso 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 Associação Brasileira de Educação Médica (ABEM), é responsável pela elaboração, expansão e aplicação do Teste de Progresso (TP) no contexto do curso de graduação em medicina.

2.2. Coleta de dados

Esse estudo utilizou como banco de questões o TP elaborado pelo Núcleo de Apoio Pedagógico Interinstitucional Sul II (NAPISUL-II), um dos grupos regionais da ABEM. A edição da prova utilizada foi a de 2023, cuja data de realização foi 18 de outubro de 2023. O documento base, do qual as questões foram transcritas, 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 Teste de Progresso 2023 para fins de pesquisa. Foi realizado contato pessoal e posterior comunicação por e-mail com a solicitação dos usos dos dados junto 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 estes dados gerais do NAPISUL -II utilizados nesta pesquisa, além de ter acesso público, são amplamente divulgados no contexto universitário.

2.3. Desenho de Estudo

De início, foram definidos quatro chatbots para análise: *ChatGPT 3.5*, *ChatGPT 4.0*, *Bing AI* e *Bard* (*Bing AI* e *Bard* posteriormente passaram a se chamar *CoPilot* e *Gemini*, respectivamente)^{31–33}, por se tratarem dos mais relevantes no momento da coleta³⁴. Os chatbots foram acessados através de suas interfaces públicas na web. Os testes de todos chatbots foram realizados 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 1 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 pré-definidas 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 o comando inicial enviado novamente. Todas as questões do TP foram encaminhadas, com exceção de questões anuladas(2), com imagens(1) ou tabelas(1) dado que, no momento da coleta, existiam diferentes limitações na capacidade de processamento e interpretação visual dos chatbots ³⁷. Após a conclusão desta etapa, todas as 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 realizado em um estudo prévio semelhante³⁸, para que a confiabilidade dos dados obtidos fossem maiores. Esta abordagem de múltiplos ensaios para cada questão é fundamental ao testar LLMs, dada a variabilidade inerente em suas respostas, garantindo uma avaliação mais robusta e representativa de seu desempenho potencial. Após isso, foi criada uma tabela através do *Google Sheets*, na qual foram evidenciadas as respostas dadas pelos chatbots ao lado do gabarito oficial do TP, para posterior avaliação de assertividade.

2.4. 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. Esta 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 décima segunda 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 na 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, ao comparar 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 décima segunda 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 estudantes da décima segunda 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 estudantes de medicina da décima segunda fase, foram indicados os intervalos em que a média de cada chatbot ficou. Além disso, foi feito 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, foi montado um gráfico que facilita a visualização da constância dos chatbots ao elucidar a quantidade de vezes que cada chatbot acertou 0, 1, 2 ou 3 vezes. Por fim, dois pesquisadores elencaram as questões de acordo com habilidades específicas da prática médica. Os grupos foram definidos de forma a contemplar mais de 10 questões cada, garantindo ao menos 30 respostas por chatbot, já que todas as questões foram respondidas 3 vezes por um mesmo chatbot. Com isso, foram definidas 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 em 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 décima segunda fase do NAPISUL-II evidenciada separadamente.

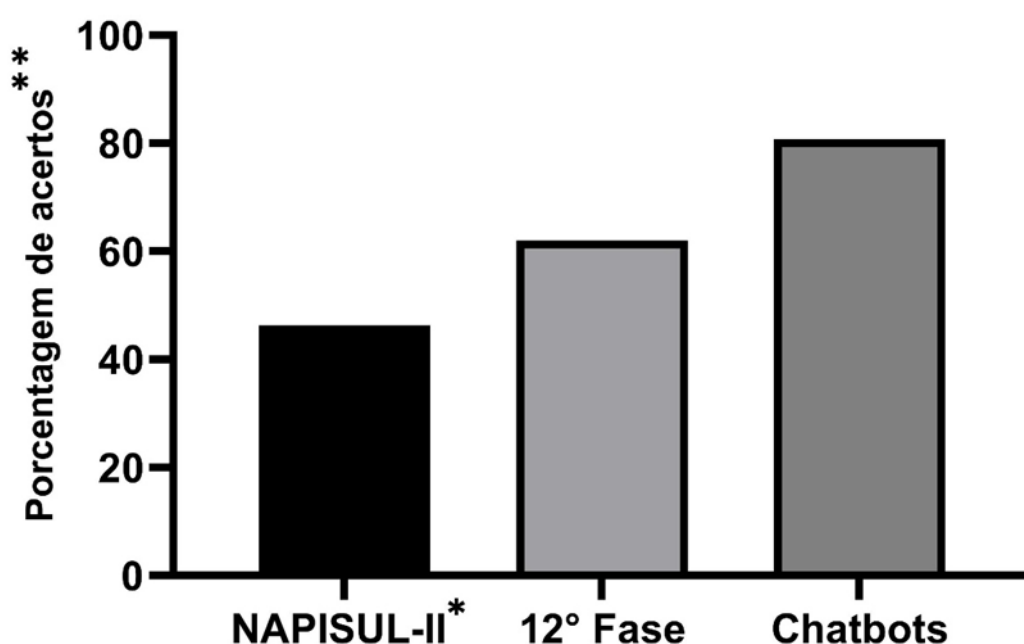
2.5. **Considerações éticas**

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 geral³⁹.

3. RESULTADOS

Os chatbots tiveram uma pontuação média de 93,73 (80,81%) acertos em 116 questões, ultrapassando os estudantes de medicina da décima segunda 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.

Figura 1. Porcentagem de acertos no teste de progresso de todos os alunos do NAPISUL-II, estudantes de 12º fase do NAPISUL-II e chatbots



*A média de acertos do NAPISUL-II corresponde a todos os estudantes do bloco NAPISUL-II.

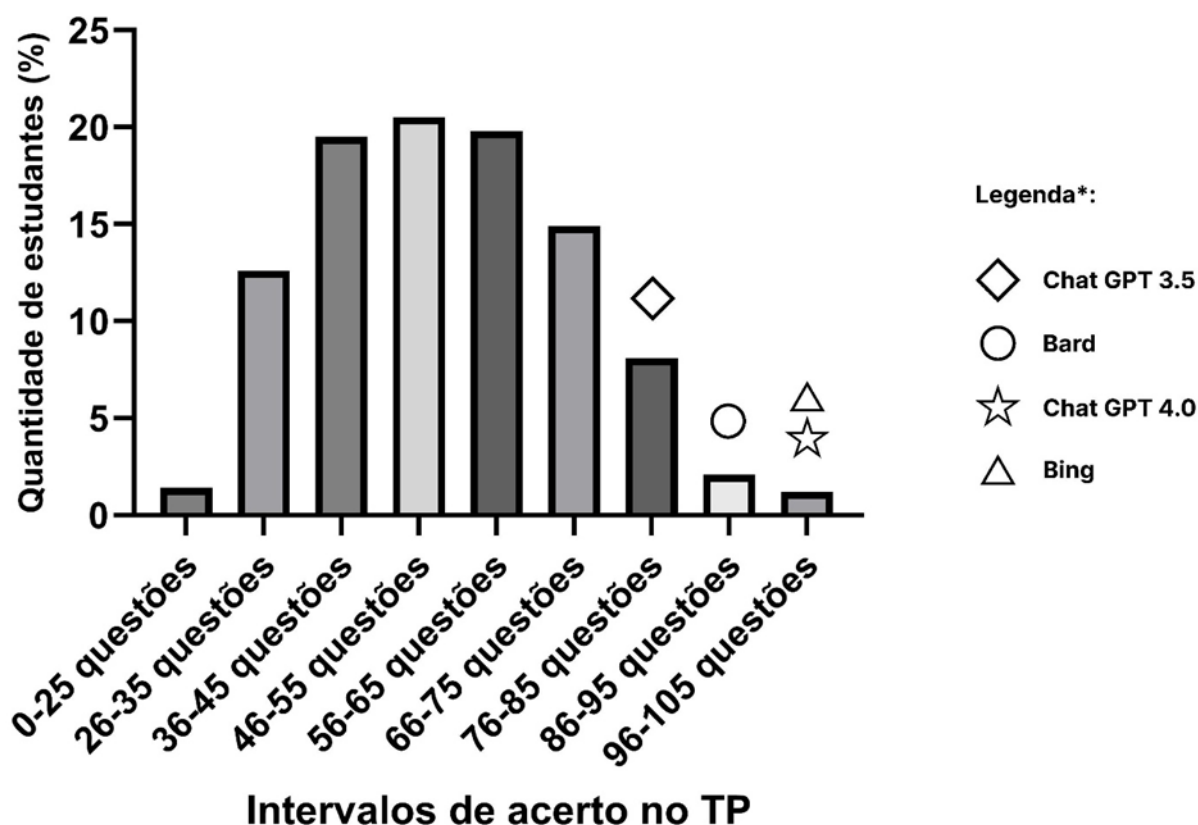
**A média do NAPISUL-II e da 12º fase foram coletadas dos resultados oficiais do TP, onde constavam 120 questões, já a média dos chatbots desconsiderou duas questões anuladas e duas questões com tabela ou imagem, sendo analisado quanto às 116 questões restantes.

Fonte: Elaborada pelos autores.

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 Chat GPT-4, que alcançou 99,6 de 116 (85,86%). Em sequência estão, respectivamente: Bard, com uma média de 89 de 116 (76,72%) e Chat GPT-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.

Figura 2. Distribuição dos Estudantes do NAPISUL-II e Chatbots por Intervalo de Acerto no Teste de Progresso



*Os símbolos representam em qual intervalo de acertos esteve o desempenho médio de cada ChatBot

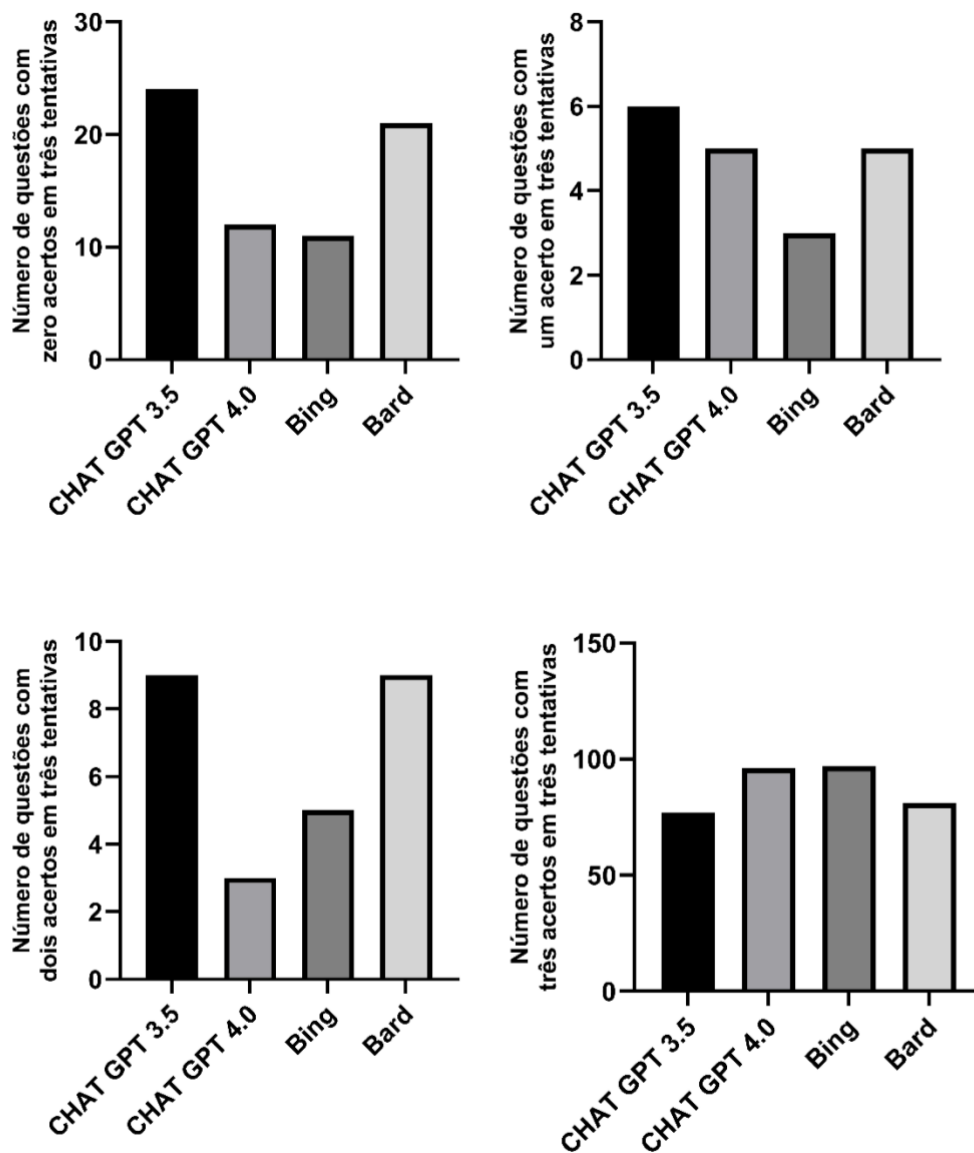
Fonte: Elaborado pelos autores

Na divisão por áreas, i.e. 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 estes. O Chat GPT-3.5, ainda que sendo o chatbot com a pior média de acertos, manteve uma média maior que os estudantes em todas as áreas. No caso do Chat GPT-4, houve destaques em clínica cirúrgica (88,33%) e clínica médica (86,67%), áreas nas quais pontuou mais que qualquer outro chatbot. O Bing, enquanto 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. Essas informações foram dispostas em uma tabela.

Tabela 1. Comparativo de percentuais de acerto entre estudantes da décima segunda fase do curso de graduação em medicina das faculdades que compõem o NAPISUL-II e chatbots no Teste de Progresso quanto às divisões por área inerente

Área do TP	12° Fase	CHAT GPT 3.5	CHAT GPT 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%

Figura 3. Número de Questões com 0, 1, 2 e 3 Acertos em 3 Tentativas por Chatbot no Teste de Progresso NAPISUL-II



Fonte: Elaborado pelos autores.

Tabela 2. Comparativo de percentuais de acerto entre estudantes da décima segunda fase do curso de graduação em medicina das faculdades que compõem o NAPISUL-II e Chatbots quanto aos principais tipos de conhecimentos médicos abordados no Teste de Progresso

Tipo de conhecimento	Número de questões*	12º Fase	Chat GPT 3.5	Chat GPT 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.

4. DISCUSSÃO

Este estudo avança significativamente em relação à literatura previamente publicada sobre o uso de LLMs no contexto do Teste de Progresso brasileiro. Enquanto o estudo de Rodrigues Alessi et al. (2024) avaliou exclusivamente o desempenho do ChatGPT 3.5 frente 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 pedagógicas²⁸.

Os resultados deste estudo destacam a notável eficácia dos chatbots de Large Language Models (LLMs) 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, com 85,86%. Esse desempenho foi superior em comparação aos estudantes do 12º período do Núcleo de Apoio Pedagógico Interinstitucional Sul II (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 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. (2024), 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 IA²⁸. No estudo de Morreel, Verhoeven e Mathysen (2024)⁴⁰, Bing AI e GPT-4 ambos 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. encontrou no Exame de Licenciamento Médico de Múltipla Escolha do Peru um resultado de 82,2% por parte do Bing, 86,7% pelo ChatGPT-4, com Bard e GPT 3.5

apresentando menor desempenho (ambos 68,8%), mais ainda superior aos estudantes (55%)³⁸.

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 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 de 12º fase obtiveram 60,50%, 67,20%, 59,40%, 62,10%, 62,50% e 60,40%. Esses resultados indicam que os chatbots não apenas superaram 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 teste de progresso, excetuando-se Clínica Médica e Clínica Cirúrgica, nas quais houve superioridade do ChatGPT4.0. Tal dado divergiu em relação Torres-Zegarra et al., em que o BING ficou atrás do ChatGPT4.0 também na área de Pediatria e Saúde Coletiva, porém apresentando mesma quantidade de acertos em Clínica Médica³⁸.

Outrossim, vale ressaltar que a demonstração de altas taxas de acerto por parte dos chatbots independente da área médica ou tipo de conhecimento avaliado é presente em estudos realizados em diversos idiomas, países e formas de cobrança de conteúdo, como pode-se observar em estudos prévios de Meo et al., Tong et al., Alijindan et al., Ebrahimian et al., Friederichs et al. e Lai et al.^{29,41-45}.

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 estudantes 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, 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, dependem

exclusivamente de seus dados de treinamento pré-existent 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 Teste de Progresso, 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, onde a obtenção da licença profissional depende do resultado e, portanto, motiva os estudantes a se empenharem ao máximo, o Teste de Progresso 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, onde a motivação é significativamente maior e o desempenho tende a ser mais representativo das habilidades dos candidatos.

Por outro lado, 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 o desempenho relacionado às diferentes áreas médicas e também 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 LLMs têm um potencial significativo para auxiliar no ensino médico. Kung et al. (2023) 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, introduzindo conceitos novos e não óbvios aos estudantes, enriquecendo a compreensão e o raciocínio clínico. De forma semelhante, Cascella et al. (2023) 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, oferecendo suporte personalizado, promovendo o desenvolvimento de habilidades clínicas e facilitando a compreensão de materiais complexos, contribuindo para um processo de aprendizagem mais autônomo e eficaz^{20,46}.

5. CONCLUSÃO

Os resultados deste estudo evidenciam que os chatbots avaliados, especialmente o Bing AI e o ChatGPT-4, 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 o aprimoramento do aprendizado e da avaliação médica.

A análise segmentada das áreas de conhecimento destacou que os chatbots, além de superar 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 onde 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 que limitações permeiam este estudo, como o viés de memória dos chatbots e a motivação reduzida dos estudantes no Teste de Progresso, o que ressalta a necessidade de uma implementação cuidadosa e criteriosa dessas tecnologias. Para que os benefícios identificados sejam plenamente aproveitados, futuras pesquisas devem focar 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.

6. REFERÊNCIAS

1. Wang X. Experiences, challenges, and prospects of National Medical Licensing Examination in China. *BMC Med Educ.* dezembro de 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 Journal for Medical Education*; 34(2):Doc25 [Internet]. 15 de maio de 2017 [citado 13 de janeiro de 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? *The Clinical Teacher.* junho de 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, Almeida 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.* junho de 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.* dezembro de 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.* dezembro de 2011;35(4):493–501.
8. Baldim TL. O teste de progresso sob a visão do discente. 2(01).
9. Martins Presta P, Pereira De Oliveira A, Rarunna Muniz Moreira ML, Otavio Ferreira Da Costa G, Peixoto RAC. Conhecimento em clínica médica: resultados de Teste de Progresso de uma instituição do Nordeste. *RI.* 28 de novembro de 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.* janeiro de 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 [Internet]. EdArXiv; 2023 jan [citado 1º de fevereiro de 2024]. Disponível em: <https://osf.io/5er8f>
12. Tareq Rasul et al. The role of ChatGPT in higher education: Benefits, challenges, and future research directions. JALT [Internet]. 11 de maio de 2023 [citado 1º de fevereiro de 2024];6(1). 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. 19 de março de 2023;11(6):887.
14. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? JALT [Internet]. 25 de janeiro de 2023 [citado 1º de fevereiro de 2024];6(1). 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. INTERSABERES. 16 de julho de 2024;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. 29 de junho de 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. 8 de fevereiro de 2023;9:e45312.
18. Shoja MM, Van De Ridder JMM, Rajput V. The Emerging Role of Generative Artificial Intelligence in Medical Education, Research, and Practice. Cureus [Internet]. 24 de junho de 2023 [citado 14 de janeiro de 2024]; Disponível em: <https://www.cureus.com/articles/158756-the-emerging-role-of-generative-artificial-intelligence-in-medical-education-research-and-practice>
19. Ramos ASM. Inteligência Artificial Generativa baseada em grandes modelos de linguagem - ferramentas de uso na pesquisa acadêmica [Internet]. 2023 maio [citado 1º de fevereiro de 2024]. Disponível em: <https://preprints.scielo.org/index.php/scielo/preprint/view/6105/version/6463>
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. Dagan A, organizador. PLOS Digit Health. 9 de fevereiro de 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*. 3 de agosto de 2023;620(7972):172–80.
22. Lee H. The rise of CHATGPT: Exploring its potential in medical education. *Anatomical Sciences Ed*. 28 de março de 2023;ase.2270.
23. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Drazen JM, Kohane IS, Leong TY, organizadores. *N Engl J Med*. 30 de março de 2023;388(13):1233–9.
24. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *Journal of the Chinese Medical Association*. julho de 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. *Aesthetic Surgery Journal*. 16 de novembro de 2023;43(12):NP1085–9.
26. Skolidis 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? *European Heart Journal - Digital Health*. 1º de junho de 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*. 1º de setembro de 2023;183(9):1028.
28. Rodrigues Alessi M, Gomes HA, Lopes De Castro M, Terumy Okamoto C. Performance of ChatGPT in Solving Questions From the Progress Test (Brazilian National Medical Exam): A Potential Artificial Intelligence Tool in Medical Practice. *Cureus [Internet]*. 19 de julho de 2024 [citado 25 de maio de 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? *Medical Education Online*. 31 de dezembro de 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*. 1º de novembro de 2023;9:e51421.
31. Chat GPT 4.0 [Internet]. Disponível em: <https://chatgpt.com/>

32. Copilot AI [Internet]. Disponível em: <https://copilot.microsoft.com/chats/naNQgAoA8PR5ctbkCv5bV>
33. Gemini AI [Internet]. Disponível em: <https://gemini.google.com/app?hl=pt-BR>
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. *European Journal of Orthodontics*. 13 de abril de 2024;cjae017.
35. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? [Internet]. arXiv; 2023 [citado 3 de julho de 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*. 1º de dezembro de 2023;31(23):1173–9.
37. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. *JAMA Ophthalmol*. 1º de junho de 2023;141(6):589.
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*. 20 de novembro de 2023;20:30.
39. Resultados Teste de Progresso [Internet]. Disponível em: 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. Banerjee I, organizador. *PLOS Digit Health*. 14 de fevereiro de 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*. 17 de julho de 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*. 19 de outubro de 2023;10:1237432.
43. Aljindan FK, Al Qurashi AA, Albalawi IAS, Alanazi AMM, Aljuhani HAM, Falah Almutairi F, 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 [Internet]. 11 de setembro de 2023 [citado 3 de julho de 2024]; Disponível em: <https://www.cureus.com/articles/172769-chatgpt-conquers-the-saudi-medical-licensing-exam-exploring-the-accuracy-of-artificial-intelligence-in-medical-knowledge-assessment-and-implications-for-modern-medical-education>

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.* dezembro de 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.* 19 de setembro de 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.* 4 de março de 2023;47(1):33.

ANEXO A – NORMAS DA REVISTA

A revista escolhida para submissão do presente artigo foi a Revista Brasileira de Educação Médica (ISSN: 1981-5271). As Diretrizes para Autores podem ser encontradas em: <https://www.scielo.br/journal/rbem/about/#instructions>- Acesso em 11/06/2025.

A **Revista Brasileira de Educação Médica** publica artigos originais, artigos de revisão, relatos de experiência, ensaios, posicionamentos, cartas ao editor e resenhas de livros sobre temas relevantes na área de educação médica. Os artigos podem ser submetidos na Língua Portuguesa ou Inglesa.

A RBEM aceita artigo preprint.

A RBEM aceita submissões contendo material que já tenha feito parte de uma tese de doutorado, dissertação de mestrado ou trabalho de conclusão de curso/iniciação científica, incluindo aquelas que foram disponibilizadas publicamente de acordo com os requisitos obrigatórios da instituição que concede a titulação, desde que obedeça a estrutura exigida pela categoria de submissão à revista.

Os manuscritos devem ser submetidos pelo sistema eletrônico [ScholarOne](#) em português ou inglês (não é permitida a alteração de idioma em nenhuma etapa após a submissão) e destinados exclusivamente à RBEM seguindo as orientações especificadas abaixo de acordo com a categoria

A contagem de palavras começa a partir da Introdução e exclui as referências. Independente da categoria, não inclui resumo/abstract, elementos figurativos, e referências bibliográficas.

Informações sobre a instituição e sobre os autores envolvidos na pesquisa que constarem no corpo do artigo devem ser sombreadas ([realce](#)) na cor preta para ocultar os dados. **Não serão aceitas substituições dessas informações por outros caracteres como XXX.**

Todo o processo editorial pode ser acompanhado pelos autores na plataforma ScholarOne, na aba referente ao Autor.

Tipos de documentos aceitos

Editorial: de responsabilidade dos editores ou de pesquisadores convidados (até 3 mil palavras). Não serão aceitos editoriais enviados espontaneamente.

Estrutura do manuscrito:

TÍTULO

DESENVOLVIMENTO LIVRE

REFERÊNCIAS

Artigo original: artigos resultantes de pesquisas originais teóricas ou empíricas (até 5 mil palavras).

Estrutura do manuscrito:

TÍTULO

RESUMO (Seções grafadas em negrito: Introdução, Objetivo, Método, Resultado, Conclusão)

PALAVRAS-CHAVE

INTRODUÇÃO

MÉTODO

RESULTADOS

DISCUSSÃO

CONCLUSÕES OU CONSIDERAÇÕES FINAIS

REFERÊNCIAS

Ensaio: artigo com análise crítica sobre um tema específico relacionado à educação médica (até 3 mil palavras).

Estrutura do manuscrito:

TÍTULO

RESUMO (Seções grafadas em negrito: Introdução, Desenvolvimento, Conclusão)

PALAVRAS-CHAVE

INTRODUÇÃO

DESENVOLVIMENTO LIVRE

CONCLUSÕES OU CONSIDERAÇÕES FINAIS

REFERÊNCIAS

Artigo de revisão: artigo baseado exclusivamente em fontes secundárias, com revisão crítica da literatura, pertinentes ao escopo da Revista (até 5 mil palavras).

Estrutura do manuscrito:

TÍTULO

RESUMO (Seções grafadas em negrito: Introdução, Objetivo, Método, Resultado, Conclusão)

PALAVRAS-CHAVE

INTRODUÇÃO

MÉTODO

RESULTADOS

DISCUSSÃO

CONCLUSÕES OU CONSIDERAÇÕES FINAIS

REFERÊNCIAS

Relato de experiência: artigo que apresente experiência inovadora na educação médica, acompanhada por reflexão teórica pertinente (até 3 mil palavras).

Estrutura do manuscrito:

TÍTULO

RESUMO (Seções grafadas em negrito: Introdução, Relato de experiência, Discussão, Conclusão)

PALAVRAS-CHAVE

INTRODUÇÃO

RELATO DE EXPERIÊNCIA

DISCUSSÃO

CONCLUSÕES OU CONSIDERAÇÕES FINAIS

REFERÊNCIAS

Carta ao editor: Correspondências de conteúdo científico contendo comentário sobre material publicado em números anteriores da Revista (até 1.200 palavras).

Estrutura do manuscrito:

TÍTULO

DESENVOLVIMENTO LIVRE

REFERÊNCIAS

Resenha: análise crítica (com reflexões e impactos para os leitores) de publicações articuladas com o ensino médico lançadas no Brasil ou no exterior (até 1.200 palavras).

Estrutura do manuscrito:

TÍTULO

DESENVOLVIMENTO LIVRE

REFERÊNCIAS

Posicionamento, Consensos e Diretrizes: os editores formulam convite a um grupo de trabalho que será responsável pela revisão aprofundada e elaboração consensuada do artigo sobre tema específico (até 8 mil palavras). Não serão aceitos posicionamentos enviados espontaneamente.

Estrutura do manuscrito:

TÍTULO

PALAVRAS-CHAVE

INTRODUÇÃO

DESENVOLVIMENTO LIVRE

CONCLUSÕES OU CONSIDERAÇÕES FINAIS

REFERÊNCIAS

Contribuição dos Autores

Preencher o formulário disponível no [link](#) após a submissão do manuscrito contendo as seguintes informações: ID da submissão, títulos em português e inglês, palavras-chave conforme o [DeCS](#), número de autores, nome, gênero, e-mail, telefone, instituição e função na mesma, número de registro [ORCID](#) e contribuição específica de cada autor para o trabalho usando as funções relevantes do [CRedit](#);

Informações sobre a existência ou não de conflito de interesses. Caso haja conflito de interesse financeiro, os autores devem informar os dados do financiamento, com o número de cadastro do projeto. Caso haja o uso de IA no manuscrito, também deve ser indicado no formulário.

Preparação do Manuscrito

- **Título:** deve conter no máximo 15 palavras e ser redigido em duas versões. Uma versão em Português e outra em Inglês.
- **Resumo:** deve conter no máximo 350 palavras e ser redigido em duas versões: inglês e português. Deve ser texto corrido e ter as seções marcadas em negrito conforme descrito na categoria do artigo.

- **Palavras-chave:** deve conter de 3 a 5 palavras extraídas dos [Descritores em Ciências da Saúde \(DeCS\)](#), para resumos em português e [Medical Subject Heading \(MeSH\)](#), para resumos em inglês.
- **Número de autores:** O número máximo de autores em todas as categorias é de seis, exceto na categoria “Posicionamentos”, cujo o número de autores será indicado pelo corpo editorial, e na categoria “Artigo original” quando o manuscrito apresentar resultados de projetos multicêntricos, quando o número de autores será de três autores por centro. Se o número de autores for superior a este, será preciso enviar uma carta com justificativa ao editor (rbem.abem@gmail.com) destacando os critérios de autoria de forma individual para cada autor. Não será aceito acréscimo de autores após o aceite do artigo.

Formato de Envio dos Artigos

Arquivo: Word, papel A4 (21 cm x 29,7 cm ou 8,3" x 11,7").

Letra: Padrão Arial 11, espaço 1,5 e margens de 2,0 cm ou 0,79" (direita, esquerda, superior e inferior).

Alinhamento: Justificado.

Parágrafos: Devem estar com recuo de 1 cm.

Títulos de seções: Colocar 1 espaço de 1,5 entre o texto do tópico anterior e o título do subsequente. Devem estar em negrito e em caixa alta.

Subtítulos: Colocar 1 espaço de 1,5 o texto do tópico anterior e o título do subsequente. Devem estar em negrito e apenas a primeira letra em maiúsculo.

Sub-subtítulos: Colocar 1 espaço de 1,5 entre o texto do tópico anterior e o título do subsequente. Devem estar em negrito, apenas a primeira letra em maiúsculo e em itálico.

Sub-sub-subtítulos: Colocar 1 espaço de 1,5 entre o texto do tópico anterior e o título do subsequente. Devem estar em negrito, apenas a primeira letra em maiúsculo, em itálico e sublinhado.

Citação até 3 linhas: Deve ser inserida no texto e estar entre aspas.

Citação com mais de 3 linhas: Deve constituir um parágrafo distinto, com recuo de 4 cm da margem esquerda, espaçamento simples, em itálico e com fonte 10.

Citação direta no corpo do artigo: Mais de 1 autor, citar o primeiro e depois adicionar et al.

Referências no corpo do artigo: Devem estar em sobrescrito, sem parênteses, antes da pontuação e sem espaço entre a palavra, o número e a pontuação (exemplos: educação médica¹. educação médica^{1,2}. educação médica¹⁻⁴. educação médica^{1,5,8-11}).

Notas de rodapé: Não serão aceitas.

Não serão publicados anexos ou arquivos suplementares.

Ativos Digitais

Representação ilustrativa: deve ter o título e a numeração na parte superior, a qual deve ter um ponto após (exemplo: Tabela 1. Título), e fonte na parte inferior. As abreviaturas, caso presentes, devem constar na primeira linha da parte inferior (Abreviaturas:). Os símbolos para explicações devem ser identificados com letras do alfabeto sobrescritas e explicados na parte inferior com fonte 10. O número máximo de arquivos é de 5.

Devem ser inseridas no corpo do artigo conforme instruções abaixo:

- Tabelas: devem conter apenas bordas horizontais.
- Figuras: devem ter boa resolução, no mínimo 300 DPI.
- Quadros: devem conter bordas horizontais e verticais em suas laterais e na separação das casas.
- Gráficos: devem conter a legenda.

Citações e Referências

Referências: a formatação segue o estilo Vancouver, conforme os Uniform Requirements for Manuscripts submitted to Biomedical Journals, publicados pelo [International Committee of Medical Journal Editors \(ICJME\)](#). As referências devem ser citadas numericamente e por ordem de aparecimento no texto. Os nomes dos periódicos devem ser abreviados de acordo com o estilo usado no Index Medicus.

[Exemplos de referências disponíveis.](#)

Documentos Suplementares

Dentre os arquivos enviados, os autores deverão incluir o [Formulário sobre Conformidade com a Ciência Aberta](#) totalmente preenchido. Este se encontra disponível na aba Instruções e formulários.

Preencher o formulário [Página de Títulos](#) online.

Declaração de financiamento

Informar fontes de apoio para o trabalho, incluindo nomes de patrocinadores, número de contrato (se houver), juntamente com explicações sobre o papel dessas fontes.