Edição Especial | Artigo Original

Recebido em: 03/05/2024 | Aceito em: 02/09/2024

Desempenho do ChatGPT e do Bard AI nas provas do exame nacional de revalidação de diplomas médicos do Brasil - Revalida

Performance of ChatGPT and BARD AI in the National Examination for the Revalidation of Medical Diplomas in Brazil

Fernanda Gabriele Fernandes Morais[1]
Orcid: https://orcid.org/0009-0007-0803-6496

Sabrine Teixeira Ferraz Grunewald[2]
Orcid https://orcid.org/0000-0003-1288-1338

Resumo

INTRODUÇÃO: O ChatGPT e o Bard AI são ferramentas de inteligência artificial construídas de forma a gerar linguagem semelhante à humana e realizar uma ampla gama de tarefas. Essas ferramentas vêm sendo estudadas quanto a inúmeras aplicações, inclusive no campo da educação médica, avaliando o desempenho em exames relevantes para o exercício profissional. OBJETIVOS: O objetivo deste estudo foi avaliar e comparar o desempenho do ChatGPT-3.5 e do Bard AI em responder às questões da prova escrita do exame nacional brasileiro para revalidação de diplomas médicos de 2023. MATERIAIS E MÉTODOS: As questões das provas objetivas foram inseridas nas ferramentas e as respostas obtidas foram comparadas com os gabaritos oficiais. As questões foram classificadas por área, cenário e complexidade. RESULTADOS: Ambas as ferramentas tiveram mais de 60% de acertos, sendo que o Bard AI foi o modelo com melhor desempenho. Não houve diferenças estatisticamente significativas no desempenho das ferramentas quando as questões foram classificadas quanto à área, cenário ou complexidade. CONCLUSÕES: É importante que os profissionais de saúde reconheçam os potenciais e as limitações dessas ferramentas, e que as pesquisas avancem para que possam ser efetivamente utilizadas na educação médica.

Palavras-chave: inteligência artificial; educação médica; desenvolvimento tecnológico.

Abstract

BACKGROUND: ChatGPT and Bard AI are artificial intelligence tools designed to generate human-like language and perform a wide range of tasks. These tools have been studied for various applications, including in the field of medical education, assessing their performance in relevant exams for professional practice. OBJECTIVE: The aim of this study was to evaluate and compare the performance of ChatGPT-3.5 and Bard AI in responding to questions from the 2023 Brazilian national exam for the revalidation of medical diplomas. METHODS: Objective exam questions were input into the tools, and the obtained responses were compared to official answer keys. Questions were categorized by area, scenario, and complexity. RESULTS: Results showed that both tools achieved over 60% accuracy, with Bard AI outperforming ChatGPT-3.5. No statistically significant differences were found in tool performance when questions were categorized by area, scenario, or complexity. CONCLUSIONS: It is crucial for healthcare professionals to recognize the potential and limitations of these tools, and further research is needed to effectively integrate them into medical education.

Keywords: artificial intelligence; education; medical; technological development.

[1] Universidade Federal de Juiz de Fora – MG – Brasil. E-mail: fernanda.gabriele@estudante.ufjf.br
[2] Universidade Federal de Juiz de Fora – MG- Brasil. E-mail: sabrine.pediatria@gmail.com

Introdução

Grandes modelos de linguagem (LLM, sigla do inglês large language models) são um tipo de tecnologia de Processamento de Linguagem Natural que têm recebido muita atenção nos últimos anos. Eles podem ser definidos como modelos de inteligência artificial (IA) treinados com grandes quantidades de dados textuais, capacitando-os a gerar linguagem semelhante à humana e realizar uma ampla gama de tarefas de processamento de linguagem[1]. Durante seu processo de treinamento, gera-se o aprendizado de padrões e estruturas da linguagem sem serem programados explicitamente, permitindo que os LLMs gerem texto coerente e contextualmente apropriado em resposta a estímulos. Alguns exemplos de LLMs disponíveis atualmente são o ChatGPT da Open AI e o Bard AI da Google.

Apesar de servirem a propósitos parecidos, as duas ferramentas de IA possuem diferenças, como o fato de o Bard AI trazer informações mais atualizadas, enquanto o Chat GPT tem seu banco de dados datados de anos anteriores de acordo com a versão. Outro ponto importante seria a capacidade de fazer buscas na internet, peculiaridade que coloca o Bard AI à frente em questão de material base para a geração de suas respostas. Por outro lado, o ChatGPT é capaz de prover texto em uma variedade de contextos maior que o Google Bard, bem como checar a presença de plágio em textos, algo que ainda não foi implementado no seu concorrente[2].

Os LLMs têm grande potencial de uso na educação médica, entretanto é importante considerar os diversos aspectos que envolvem o uso dessas ferramentas, avaliando antagonicamente o impacto dos fatores positivos e negativos. É inegável que esses modelos oferecem acesso a informações médicas abundantes, além da possibilidade de serem utilizados para tradução de textos, aumentando a diversidade de materiais que podem ser consultados na educação médica. A capacidade de responder a perguntas complexas e fornecer insights aprofundados em diversas áreas da medicina é uma vantagem significativa, auxiliando tanto alunos quanto professores. Além disso, a geração automatizada de materiais educacionais pode economizar tempo para os educadores e alunos, permitindo uma abordagem mais eficiente no processo de ensino[3].

No entanto, é crucial considerar também que os algoritmos dos LLMs podem refletir vieses ou decisões éticas questionáveis, requerendo supervisão humana para garantir a integridade e ética na educação médica. A dependência excessiva desses modelos traz o risco de aprendizado de informações incorretas ou desatualizadas, destacando a importância de verificar que os LLM estão com treinamento adequado e atualização constante, além do cuidado de se consultar informações confiáveis em diferentes fontes. Outra preocupação a se considerar é a possibilidade de que o uso excessivo dos LLMs suprima as habilidades naturais de resolução de problemas, criatividade e habilidades de aprendizado dos estudantes[2].

Portanto, ao explorar o potencial dos LLMs na educação médica, é essencial equilibrar os benefícios oferecidos com uma abordagem cuidadosa para mitigar os desafios apresentados por essas tecnologias inovadoras. Uma das linhas de pesquisa envolvendo os LLMs na educação médica é a avaliação de seu desempenho em processos avaliativos formais, como exames para admissão em programas de residência médica ou para validação do exercício profissional. Nesse sentido, o ChatGPT foi avaliado em algumas pesquisas, apresentando desempenho satisfatório[4,5]. Pesquisas com o Bard AI, entretanto, ainda são escassas.

O Revalida, ou Exame Nacional de Revalidação de Diplomas Médicos Expedidos por Instituição de Educação Superior Estrangeira, é um processo que facilita a revalidação de diplomas de médicos formados no exterior que desejam atuar no Brasil. Destinado tanto a estrangeiros formados em medicina fora do país como a brasileiros graduados em outros países, o exame consiste em duas etapas eliminatórias realizadas em momentos distintos: provas escritas e avaliação de habilidades clínicas. O exame tem como base a demonstração de conhecimentos, habilidades e competências essenciais para o exercício da medicina. A aprovação em ambas as etapas atesta a competência técnica (teórica e prática) do médico graduado para o exercício profissional[6]. O desempenho do ChatGPT na edição de 2022 do Revalida foi analisado em uma pesquisa prévia[7].

Diante do exposto, este estudo tem como objetivo avaliar e comparar o desempenho do ChatGPT-3.5 e do Bard AI em responder às questões da prova escrita do Revalida, edições de 2023 de acordo com a área, cenário e complexidade específicos de cada questão.

Materiais e Métodos

Seleção de provas e questões

Foram selecionadas as duas provas objetivas mais recentes do Revalida, disponibilizadas na íntegra no website do Inep, bem como seus gabaritos oficiais: a Prova Objetiva 1 de 2023 e a Prova Objetiva 2 de 2023. Cada prova é composta de 100 questões de múltipla escolha, com quatro opções e uma única resposta correta. As provas discursivas das respectivas edições não foram analisadas.

É importante destacar que a nota de corte para aprovação é variável para cada edição do Revalida, e inclui também o desempenho na prova discursiva, que é realizada pelos candidatos no mesmo dia da prova objetiva. A nota de corte é calculada de acordo com o nível de dificuldade da edição, avaliado pela Comissão de Avaliação de Itens. Para a primeira edição de 2023, a nota de corte foi de cerca de 64%, e para a segunda edição de 2023, a nota de corte foi de cerca de 67% da pontuação total.

Obtenção de respostas pelas ferramentas de IA

Para todas as questões da prova 1, as respostas do ChatGPT e do Bard AI foram obtidas em 18 de novembro de 2023. Para todas as questões da prova 2, as respostas de ambas as ferramentas foram obtidas em 19 de novembro de 2023. Cada questão era inserida individualmente na caixa de texto da ferramenta, e a resposta gerada era registrada. Todas as questões foram inseridas no idioma português brasileiro.

Categorização das questões

As questões das provas 1 e 2 foram categorizadas quanto a três aspectos: área (Clínica Geral, Cirurgia, Ginecologia e Obstetrícia, Pediatria e Medicina Preventiva); cenário (unidade de atenção primária, ambulatório de especialidades, urgência e emergência, hospitalar, e não informado ou não se aplica); e complexidade.

A avaliação da complexidade foi realizada de acordo com o conceito da taxonomia de Bloom, relacionando-se com o grau de raciocínio clínico e abstração necessários para responder à questão. Dessa forma, questões cujo objetivo é avaliar a capacidade de lembrar informações são classificadas como de conhecimento; questões de compreensão envolvem a habilidade de dar significado aos conteúdos; já questões que priorizam a habilidade de usar informações em situações concretas são consideradas como de aplicação; questões de análise requerem que o conteúdo seja subdividido para compreensão; questões de síntese envolvem a combinação de partes do conteúdo para formação do todo; e, por fim, questões de

avaliação são as que exercitam a capacidade de julgar conceitos e situações a partir de critérios definidos[8,9].

As questões foram avaliadas quanto aos aspectos de área, cenário e complexidade por dois pesquisadores individualmente, e as opiniões foram comparadas. Para as questões em que havia divergência inicial, foi obtido consenso entre os pesquisadores. A complexidade das questões foi dicotomizada, de forma que questões de conhecimento, compreensão e aplicação foram consideradas de baixa complexidade (predomínio de raciocínio teórico), enquanto as questões de análise, síntese e avaliação foram consideradas de alta complexidade (predomínio de raciocínio clínico)[8,10].

Análise estatística

Os dados coletados foram organizados em uma planilha do programa Microsoft Excel e os testes estatísticos, realizados no programa Statistical Package for the Social Sciences (SPSS), versão 22.0. A análise descritiva foi realizada com cálculo frequências, média e desvio padrão para as variáveis estudadas. O teste do Qui Quadrado de Pearson foi utilizado para avaliar a existência de diferença estatística entre as variáveis categóricas, adotando-se significância de $p < 0,05$.

Resultados

A prova do Revalida é composta por 100 questões. Na edição de 2023, foram anuladas, pela comissão organizadora, sete questões da prova 1 e nove questões da prova 2, que foram excluídas da análise. Assim, os resultados apresentados referem-se às 93 questões válidas da prova 1, e as 91 questões válidas da prova 2.

Do total de questões, o ChatGPT acertou 61/93 (65,6%) da prova 1, enquanto o Bard AI acertou 70/93 (75,3%). Para a prova 2, o ChatGPT acertou 54/91 (59,3%) questões, enquanto o Bard AI acertou 61/91 (67,0%).

A prova 1 apresentava quatro questões com imagens ou tabelas, enquanto a prova 2 apresentava 12 questões com imagens ou tabelas. A presença de imagens é um limitador ao desempenho de ambas as ferramentas, pois elas não foram configuradas para reconhecê-las e interpretá-las. Assim, para essas questões, as ferramentas de IA tiveram acesso apenas ao texto das mesmas. As duas ferramentas acertaram 2/4 (50%) questões com imagens na prova 1. Na prova 2, o ChatGPT acertou 6/12 (50%) das questões com imagens e o Bard AI acertou 8/12 (75%) das questões com imagens.

Para as demais análises apresentadas a seguir, as questões com imagens foram excluídas, resultando em 89 questões válidas e sem imagens na prova 1, e 79 questões válidas sem imagens na prova 2.

O desempenho de ambas as ferramentas foi superior a 60% de acertos em ambas as provas, e o Bard AI apresentou um maior número de acertos em comparação com o ChatGPT. Ao avaliar o número de acertos de acordo com a área da questão, os piores desempenhos do ChatGPT foram para a área de Ginecologia e Obstetrícia, com 50% de acertos na prova 1, e 35,3% na prova 2. Para o Bard AI, os piores desempenhos foram em Cirurgia na prova 1 (56,3% de acertos) e Ginecologia e Obstetrícia na prova 2 (52,9% de acertos). A Tabela 1 traz detalhadamente o número de acertos de cada uma das IAs em ambas as provas.

Tabela 1 - Número absoluto e relativo de acertos pelas duas ferramentas de inteligência artificial nas duas provas do Revalida 2023, divididos por área da questão (apenas questões válidas e sem imagens).

	ChatGPT n (%)	**Bard AI n (%)**
Prova 1 (n° de questões)		
Clínica geral (18)	13 (72,2)	15 (83,3)
Cirurgia (16)	11 (68,8)	9 (56,3)
Pediatria (19)	12 (63,2)	15 (79,0)
Ginecologia e Obstetrícia (16)	8 (50,0)	12 (75,0)
Medicina Preventiva (20)	15 (75,0)	17 (85,0)
Total (89)	59 (66,3)	68 (76,4)
Prova 2 (n° de questões)		
Clínica geral (18)	13 (72,2)	15 (83,3)
Cirurgia (16)	10 (62,5)	10 (62,5)
Pediatria (11)	6 (54,6)	6 (54,6)
Ginecologia e Obstetrícia (17)	6 (35,3)	9 (52,9)
Medicina Preventiva (17)	13 (76,5)	13 (76,5)
Total (79)	48 (60,8)	53 (67,1)

Fonte: Dados da pesquisa

Na avaliação dos acertos das ferramentas de IA, de acordo com o cenário da questão, o cenário de Urgência e Emergência representou o pior desempenho para o ChatGPT em ambas as provas, enquanto o cenário Hospitalar foi aquele com menor número de acertos para o Bard AI. No entanto, a diferença de acertos entre os cenários das questões não alcançou significância estatística para nenhuma das provas. Os detalhes sobre a frequência de acertos por cenários estão disponíveis na Tabela 2.

Tabela 2 - Número absoluto e relativo de acertos pelas duas ferramentas de inteligência artificial nas duas provas do Revalida 2023, divididos por cenário da questão.

	ChatGPT n (%)	**Bard AI n (%)**
Prova 1 (n° de questões)		
Atenção primária (25)	19 (76,0)	20 (80,0)
Urgência e Emergência (21)	12 (57,1)	15 (71,4)
Ambulatórios de especialidades (16)	10 (62,5)	13 (81,3)
Hospitalar (9)	6 (66,7)	4 (44,4)
Não informado/Não se aplica (18)	12 (66,7)	16 (88,9)
Total (89)	59 (66,3)	68 (76,4)
Valor de *p*	0,75	0,12
Prova 2 (n° de questões)		
Atenção primária (30)	18 (60,0)	19 (63,3)
Urgência e Emergência (17)	9 (52,9)	10 (58,8)
Ambulatórios de especialidades (9)	5 (55,6)	9 (100,0)
Hospitalar (10)	7 (70,0)	5 (50,0)
Não informado/Não se aplica (13)	9 (69,2)	10 (76,9)
Total (79)	48 (60,8)	53 (67,1)
Valor de *p*	0,86	0,14

Fonte: Dados da pesquisa

O número de acertos das ferramentas de IA também não diferiu de forma estatisticamente significativa quando foram agrupados de acordo com a complexidade das questões de cada prova, avaliada à luz dos conceitos da Taxonomia de Bloom (Tabela 3).

Tabela 3 - Número absoluto e relativo de acertos pelas duas ferramentas de inteligência artificial nas duas provas do Revalida 2023, divididos por complexidade da questão, de acordo com a taxonomia de Bloom.

	ChatGPT n (%)	**Bard AI n (%)**
Prova 1 (n° de questões)		
Baixa complexidade (36)	24 (66,7)	30 (83,3)
Alta complexidade (53)	35 (66,0)	38 (71,7)
Valor de *p*	1,00	0,21
Prova 2 (n° de questões)		
Baixa complexidade (20)	12 (60,0)	12 (60,0)
Alta complexidade (59)	36 (61,0)	41 (69,5)
Valor de *p*	0,94	0,46

Fonte: Dados da pesquisa

Discussão

As ferramentas de IA ChatGPT e Bard AI foram comparadas quanto ao número de acertos nas provas de duas edições de 2023 do Revalida, obtendo mais de 60% de acertos, sendo que o Bard AI foi o modelo com melhor desempenho em ambas as provas. Não houve diferenças estatisticamente significativas no desempenho das ferramentas quando as questões foram classificadas quanto à área, cenário ou nível de complexidade.

Os resultados da acurácia do ChatGPT foram avaliados para diversos exames da área médica. Um estudo avaliou o desempenho do ChatGPT no Exame Médico para Licenciamento dos Estados Unidos (USMLE), com percentual de acertos superior a 60% nos diversos conjuntos de questões analisados, com demonstração de raciocínio em suas respostas dialogadas[4]. Uma outra pesquisa comparou o desempenho das versões 3.5 e 4.0 do ChatGPT no Exame Médico para Licenciamento do Japão, mostrando que apenas a versão 4.0, que é de utilização paga, alcançou a nota mínima para aprovação, com melhor desempenho sobretudo nas questões consideradas mais difíceis[5]. Resultados semelhantes também foram encontrados para o exame de licenciamento polonês[11].

Já, um estudo brasileiro avaliou o comportamento do ChatGPT (versão 4.0) na segunda edição do Revalida de 2022, que apresentou um percentual de acertos de 87,7%, sem diferença estatística entre as áreas das questões[7]. O desempenho superior verificado nessa pesquisa, em comparação com nossos resultados, pode ser justificado em parte pelo uso da versão paga da ferramenta.

Pesquisas avaliando o desempenho do Bard AI ou comparando-o com o do ChatGPT também começam a surgir na literatura. Em um estudo que avaliou as respostas das ferramentas a vinhetas com casos clínicos sobre fisiologia, o ChatGPT teve melhor desempenho que o Bard AI[12], e resultados semelhantes foram encontrados pela mesma equipe, em uma pesquisa na qual foi avaliado o desempenho dos modelos de IA frente a questões de hematologia[13]. Já, um estudo que comparou

a capacidade das ferramentas em gerar questões de múltipla escolha sobre fisiologia, o ChatGPT gerou mais questões válidas, porém com um nível de dificuldade inferior às questões geradas pelo Bard AI[14].

Segundo informações do Inep, a taxa de aprovação dos candidatos na primeira etapa do Revalida foi de 9,12% para a primeira edição de 2022; 12,71% para a segunda edição de 2022; 13,42% para a primeira edição de 2023; e 12,74% para a segunda edição de 2023[15]. Vários fatores são apontados como potenciais explicações para essas baixas taxas de aprovação, incluindo qualidade da instrução médica, preparo dos candidatos, efetividade dos métodos de avaliação e barreias econômicas e culturais[7]. As notas obtidas pelas ferramentas de IA no presente estudo seriam suficientes para a aprovação na prova 2, mas apenas o Bard AI obteve a pontuação mínima para a prova 1. Esse desempenho dos modelos de IA sugere seu potencial em superar o desempenho humano em tarefas de alta complexidade em campos altamente especializados[5].

Os modelos de IA vêm tendo seu potencial explorado em diversas estratégias de educação médica, como para o ensino de pequenos grupos e aprendizado por pares, auxiliando os estudantes em seu estudo independente[5]. Um exemplo aplicável ao contexto da prova do Revalida seria a discussão pelos estudantes das respostas apresentadas pela ferramenta de IA, buscando compreender as justificativas para essas respostas, e, no caso de erros, qual a resposta certa e por quê.

O presente estudo apresenta algumas limitações. Primeiro, as ferramentas de IA foram utilizadas apenas em suas versões gratuitas, que podem ter desempenho inferior às versões pagas. Segundo, o desempenho das ferramentas tende a se modificar ao longo do tempo, geralmente no sentido da melhoria, portanto os resultados obtidos podem não ser replicáveis no futuro. Terceiro, foi necessária a exclusão de questões com imagens e tabelas, pois no momento da pesquisa, ambos os modelos não eram capazes de compreendê-las. Isso pode mudar com novas atualizações.

O avanço dos modelos de IA é notável, mas ainda é necessário a realização de mais pesquisas que verifiquem sua acurácia e relevância como ferramentas de educação para áreas de alta complexidade como as ciências médicas. Considerando o desempenho dos modelos de IA avaliados no presente estudo, é importante para os profissionais de saúde reconhecerem os potenciais e as limitações dessas ferramentas. Além disso, apesar do número crescente de pesquisas sobre o uso dos LLMs na educação médica, ainda é preciso se verificar, através de estudos, a real precisão e efetividade dessas ferramentas em diferentes tarefas.

Conclusão

Nesta pesquisa, o desempenho das ferramentas de inteligência artificial na resposta a questões da prova do exame nacional para revalidação de diplomas médicos foi surpreendente. Na medida em que essas ferramentas são desenvolvidas e aprimoradas, seu uso pode ser explorado em contextos variados, inclusive na educação médica, o que deve ser pautado por pesquisas que avaliem sua eficácia e factibilidade.

Referências Bibliográficas

1. Schwenk, H. Continuous Space Language Models. Computer Speech & Language, vol. 21, n. 3, Jul. 2007, p. 492–518.

2. Singh, S. K.; Kumar, S.; Mehra, P. S. Chat GPT & Google Bard AI: a Review. In: International Conference on Iot, Communication and Automation Technology (ICICAT). 23 jun. 2023, doi: 10.1109/ICICAT57735.2023.10263706. Acessado em 01 fev. 2024.

3. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare, vol. 11, no. 6, 19 Mar. 2023, p. 887, doi: 10.3390/healthcare11060887. Acessado em 01 fev. 2024

4. Gilson, A., *et al.* How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Medical Education, vol. 9, Fev. 2023, e45312, doi: 10.2196/45312. Acessado em 01 fev. 2024.

5. Takagi, S., et al. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Medical Education, vol. 9, Jun. 2023, e48002, doi: 10.2196/48002. Acessado em 01 fev. 2024.

6. Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Painel Revalida. Brasília: Inep, 2022. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/inep-data/painel-revalida. Acessado em 01 fev. 2024.

7. Gobira, M., *et al*. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. Revista da Associação Médica Brasileira, vol. 69, n. 10, 2023, p. 1-5, doi:10.1590/1806-9282.20230848. Acessado em 01 fev. 2024.

8. Ferraz, A.P.C.M.; Belhot, R.V. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. Gestão & Produção, vol. 17, n. 4, 2010, p. 421-31.

9. Aragão J.C.S., *et al*. Evaluation of Residency Admission Exams. Revista Brasileira de Educação Médica, vol. 42, n. 2, Abr. 2018, p. 26-33, doi:10.1590/1981-52712015v421n2RB20170016. Acessado em 01 fev. 2024.

10. Casiraghi, B., *et al.* Avaliação de questões de prova do Revalida no Brasil. 2019. XV Congreso Internacional Gallego-Portugués De Psicopedagogía. Disponível em: https://ruc.udc.es/dspace/handle/2183/23486. Acessado em 01 fev. 2024.

11. Wójcik, S., *et al*. Reshaping medical education: Performance of ChatGPT on a PES medical examination. Cardiology Journal, Out. 2023, doi: 10.5603/cj.97517. Acessado em 01 fev. 2024.

12. Dhanvijay, A.K.D., et al. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. Cureus. vol. 15, n. 8, Ago. 2023, p. e42972.

13. Kumari, A., *et al*. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. Cureus., vol. 15, n. 8, Ago. 2023, p. e43861.

14. Agarwal, M.; Sharma, P.; Goswami, A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. vol. 15, n. 6, Jun. 2023, p. e40977.

15. Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Exame Nacional de Revalidação de Diplomas Médicos Expedidos por Instituição de Educação Superior Estrangeira (Revalida). Brasília: Inep, 2024. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/revalida. Acessado em 01 fev. 2024.

Como citar este artigo:

Morais FGF, Grunewald STF. Desempenho do ChatGPT e do Bard AI nas provas do exame nacional de revalidação de diplomas médicos do Brasil - Revalida. Rev. Aten. Saúde. 2024; e20249478(22). doi https//doi.org/10.13037/ras.vol22.e20249478