

Desempenho do ChatGPT nas questões da avaliação anual de residentes do Colégio Brasileiro de Radiologia

Performance of ChatGPT on questions from the Brazilian College of Radiology annual resident evaluation test

Cleverson Alex Leitão^{1,a}, Gabriel Lucca de Oliveira Salvador^{1,b}, Leda Maria Rabelo^{1,c}, Dante Luiz Escuissato^{1,d}

1. Universidade Federal do Paraná (UFPR), Curitiba, PR, Brasil.

Correspondência: Dr. Cleverson Alex Leitão. Universidade Federal do Paraná. Rua General Carneiro, 181, Alto da Glória. Curitiba, PR, Brasil, 80060-900. E-mail: cleverleitao@gmail.com.

a. <https://orcid.org/0000-0003-0463-0643>; b. <https://orcid.org/0000-0001-9776-6851>; c. <https://orcid.org/0000-0001-8733-0755>;

d. <https://orcid.org/0000-0002-8978-4897>.

Submetido em 13/7/2023. Revisado em 29/8/2023. Aceito em 15/9/2023.

Como citar este artigo:

Leitão CA, Salvador GLO, Rabelo LM, Escuissato DL. Desempenho do ChatGPT nas questões da avaliação anual de residentes do Colégio Brasileiro de Radiologia. Radiol Bras. 2024;57:e20230083.

Resumo Objetivo: Testar o desempenho do ChatGPT em questões de radiologia formuladas pelo Colégio Brasileiro de Radiologia (CBR), avaliando seus erros e acertos.

Materiais e Métodos: 165 questões da avaliação anual dos residentes do CBR (2018, 2019 e 2022) foram apresentadas ao ChatGPT. Elas foram divididas, para análise estatística, em questões que avaliavam habilidades cognitivas de ordem superior ou inferior e de acordo com a subespecialidade, o tipo da questão (descrição de um achado clínico ou sinal, manejo clínico de um doente, aplicação de um conceito, cálculo ou classificação dos achados descritos, associação entre doenças ou anatomia) e o ano da residência (R1, R2 ou R3).

Resultados: O ChatGPT acertou 53,3% das questões (88/165). Houve diferença estatística entre o desempenho em questões de ordem cognitiva inferior (64,4%; 38/59) e superior (47,2%; 50/106) ($p = 0,01$). Houve maior índice de acertos em física (90,0%; 18/20) do que em questões clínicas (48,3%; 70/145) ($p = 0,02$). Não houve diferença significativa de desempenho entre subespecialidades ou ano de residência ($p > 0,05$).

Conclusão: Mesmo sem treinamento dedicado a essa área, o ChatGPT apresenta desempenho razoável, mas ainda insuficiente para aprovação, em questões de radiologia formuladas pelo CBR.

Unitermos: Inteligência artificial; Radiologia; Questões de prova; Diagnóstico por imagem.

Abstract Objective: To test the performance of ChatGPT on radiology questions formulated by the Colégio Brasileiro de Radiologia (CBR, Brazilian College of Radiology), evaluating its failures and successes.

Materials and Methods: 165 questions from the CBR annual resident assessment (2018, 2019, and 2022) were presented to ChatGPT. For statistical analysis, the questions were divided by the type of cognitive skills assessed (lower or higher order), by topic (physics or clinical), by subspecialty, by style (description of a clinical finding or sign, clinical management of a case, application of a concept, calculation/classification of findings, correlations between diseases, or anatomy), and by target academic year (all, second/third year, or third year only).

Results: ChatGPT answered 88 (53.3%) of the questions correctly. It performed significantly better on the questions assessing lower-order cognitive skills than on those assessing higher-order cognitive skills, providing the correct answer on 38 (64.4%) of 59 questions and on only 50 (47.2%) of 106 questions, respectively ($p = 0.01$). The accuracy rate was significantly higher for physics questions than for clinical questions, correct answers being provided for 18 (90.0%) of 20 physics questions and for 70 (48.3%) of 145 clinical questions ($p = 0.02$). There was no significant difference in performance among the subspecialties or among the academic years ($p > 0.05$).

Conclusion: Even without dedicated training in this field, ChatGPT demonstrates reasonable performance, albeit still insufficient for approval, on radiology questions formulated by the CBR.

Keywords: Artificial intelligence; Radiology; Examination questions; Diagnostic imaging.

INTRODUÇÃO

Inteligência artificial (IA) é o nome geral dado a métodos de computação que simulam o padrão de aprendizado do intelecto humano⁽¹⁾. Os rápidos avanços obtidos recentemente nesse campo do conhecimento têm suscitado questionamentos acerca de como ela impactará as mais

diversas profissões no futuro, inclusive a medicina. Dentre os modelos de IA já existentes, o *Chat Generative Pre-trained Transformer* (ChatGPT) tem ganhado destaque não apenas na literatura científica⁽²⁻⁴⁾, mas também na mídia comum⁽⁵⁾. Trata-se de uma ferramenta de IA baseada em relações entre algoritmos de IA com a linguagem humana

denominada *natural language processing* disponível publicamente desde 30 de novembro de 2022⁽⁶⁾. Seu algoritmo atual é o GPT-3.5, um modelo de linguagem grande treinado com mais de 45 terabytes de dados textuais. Por meio de redes neurais, tais dados permitem que a ferramenta seja capaz de analisar e gerar textos similares aos escritos por humanos⁽⁷⁾. Embora não tenha sido treinado especificamente para uso médico, estudos já têm demonstrado seu promissor papel tanto na prática médica⁽⁸⁾ quanto na escrita acadêmica em medicina⁽⁹⁾.

Como forma de avaliar o conhecimento do ChatGPT acerca de temas médicos, a ferramenta já teve seu desempenho testado em exames acadêmicos que avaliam estudantes reais, como a prova dos Estados Unidos para obtenção da licença médica⁽¹⁰⁾ e questões para a obtenção de títulos de especialista em radiologia no Canadá e nos Estados Unidos⁽⁷⁾ e em medicina de família em Taiwan⁽¹¹⁾, com resultados que mostram um desempenho, em geral, próximo ao necessário para aprovação.

No contexto específico da radiologia, a IA vem sendo usada sobretudo como auxiliar na interpretação de imagens, porém modelos de linguagem como o ChatGPT também têm mostrado potencial como auxiliar na redação dos relatórios radiológicos⁽¹²⁾ e na tomada de decisões clínicas⁽⁴⁾. Compreender melhor o desempenho da IA frente a problemas encontrados na prática diária do radiologista pode auxiliar a entender como ela influenciará o futuro da profissão. Com esse objetivo em vista, este trabalho busca avaliar o desempenho do ChatGPT em questões elaboradas pelo Colégio Brasileiro de Radiologia (CBR) para a avaliação anual dos residentes em radiologia e diagnóstico por imagem, analisando suas respostas, e quais os acertos e erros da plataforma até este momento.

MATERIAIS E MÉTODOS

Estudo analítico prospectivo realizado entre os dias 24 de maio e 3 de junho de 2023, que não envolveu seres humanos ou dados de pacientes, sendo dispensada a aprovação por comitê de ética institucional.

Questões da avaliação anual dos residentes em radiologia

Foram selecionadas 165 questões das provas anuais de avaliação de residentes em radiologia e diagnóstico por imagem aplicadas pelo CBR nos anos de 2018, 2019 e 2022, que se encontram disponíveis *on-line* para acesso público no site do CBR⁽¹³⁾ e cujo uso foi autorizado pela sua Comissão de Admissão e Titulação. Todas as questões eram do tipo “múltipla escolha”, com apenas uma alternativa correta e quatro alternativas falsas. Foram excluídas questões com imagens, pois o ChatGPT ainda não possui a capacidade de interpretá-las. Elas foram divididas de acordo com seu tema em questões de física (20 questões) e clínicas (145 questões), essas representando os principais campos de conhecimento e subespecialidades da radiologia: abdome (20

questões), tórax (15 questões), mama (15 questões), neurorradiologia (15 questões), pediatria (15 questões), musculoesquelético (15 questões), meios de contraste (15 questões), ultrassonografia (15 questões), ginecologia e obstetrícia (10 questões) e miscelânea (PET/CT, densitometria, Doppler e segurança do paciente – 10 questões).

Posteriormente, as questões foram subdivididas de acordo com os princípios da taxonomia de Bloom em questões que avaliam habilidades cognitivas de ordem inferior (relembrar uma ideia, memorizar um conceito) e questões que avaliam habilidades cognitivas de ordem superior (avaliar, analisar, sintetizar o conhecimento obtido)⁽¹⁴⁾. Estas foram novamente divididas de acordo com seu estilo em seis subcategorias: descrição de um achado clínico ou sinal; manejo clínico de um doente; aplicação de um conceito; cálculo ou classificação dos achados descritos; associação entre doenças; anatomia. Todas as questões foram classificadas independentemente pelos autores do estudo e, nos casos de desavença, uma classificação final foi obtida por consenso.

Por fim, as questões foram divididas entre questões que o CBR aplicou para os residentes dos três anos (n = 92), questões voltadas para residentes do segundo e terceiro anos (n = 34) e questões voltadas para os residentes do terceiro ano (n = 39).

ChatGPT

Foi utilizada a versão mais recente do ChatGPT disponível (24 de maio de 2023; OpenAI). Apesar de essa ferramenta ter sido treinada com mais de 45 terabytes de dados em formato de texto, advindos de páginas da internet, livros e artigos científicos, eles não foram fornecidos especificamente para atender às necessidades do médico radiologista. O ChatGPT não realiza buscas na internet, respondendo às perguntas utilizando apenas a sua base de dados.

Coleta e análise dos dados

As questões e suas alternativas foram apresentadas ao ChatGPT sequencialmente, uma a uma, tal como formuladas pelo CBR, sem o fornecimento de um *pré-prompt* específico, e suas respostas foram salvas em um arquivo de texto para a análise posterior dos pesquisadores. Para as questões respondidas incorretamente, foi imediatamente fornecida uma devolutiva, explicando o erro e qual a resposta correta, a fim de analisar o comportamento da plataforma frente a essa correção. Além da análise quantitativa do número de acertos e erros, os pesquisadores realizam uma análise qualitativa em grupo, obtendo um consenso para os comentários a respeito das respostas obtidas.

Análise estatística

Para análise do índice de acertos, foi calculada a razão do número de acertos sobre o número total de questões para todas as categorias (total de questões, questões de

alta e baixa ordem e subtipos de questões conforme atrás descrito) e exibida a porcentagem final dessa razão.

A comparação entre o índice de acertos nos diferentes grupos (baixa ordem × alta ordem; física × clínica; tipo de questão) utilizou os testes exato de Fisher e qui-quadrado. A análise entre subgrupos de questões (por tema e ano da questão) utilizou o teste de análise de variância. O programa utilizado foi o Stata versão 16.0 (Stata Corp LP, College Station, TX, EUA) e o pós-processamento foi realizado no Microsoft Excel 365, com seu pacote de análise de dados. Consideraram-se como estatisticamente significativos valores de $p < 0,05$.

RESULTADOS

Resultado geral

O ChatGPT acertou 53% das questões que lhe foram apresentadas (88/165), abaixo da nota de aprovação do CBR (70%). A Tabela 1 mostra o desempenho de acordo com o tipo e o tema da questão.

Desempenho por tipo de questão

O desempenho em questões que avaliam habilidades cognitivas de ordem inferior (64,4%; 38/59) foi superior ao visto nas questões que avaliam habilidades cognitivas de ordem superior (47,2%; 50/106) ($p = 0,01$). A Figura 1 mostra um exemplo de acerto de uma questão que avalia habilidade de ordem inferior e a Figura 2 mostra um exemplo de acerto de ordem superior.

Tabela 1—Desempenho do ChatGPT de acordo com o tipo e o tema da questão.

Característica da questão	Número de questões	Acertos n (%)	P
Tipo			
Ordem inferior	59	38 (64,4)	
Ordem superior	106	50 (47,2)	
Descrição de achados	42	22 (52,4)	0,81*
Manejo clínico	22	12 (54,5)	0,72*
Aplicar conceito	57	38 (66,7)	0,67*
Cálculo/classificação	8	3 (37,5)	0,92*
Associar doenças	26	11 (42,3)	0,63*
Anatomia	10	2 (20,0)	0,58*
Tema			
Física	20	18 (90,0)	0,02*
Clínica	145	68 (46,8)	0,41*
Abdome	20	13 (65,0)	0,62†
Tórax	15	9 (60,0)	0,56†
Neurorradiologia	15	5 (33,3)	0,76†
Musculoesquelético	15	8 (53,3)	0,87†
Mama	15	7 (46,7)	0,61†
Meios de contraste	15	9 (60,0)	0,94†
Ultrassonografia	15	3 (20,0)	0,78†
Pediatria	15	10 (66,7)	0,93†
Ginecologia e obstetrícia	10	2 (20,0)	0,72†
Miscelânea	10	4 (40,0)	0,65†
Total	165	88 (53,3)	0,01*

* Teste exato de Fisher. † Análise de variância.

C Assinale a alternativa que representa a etiologia da atenuação em mosaico em uma TCAR de tórax.

A) Preenchimento alveolar parcial com espessamento intersticial.

B) Espessamento de paredes bronquiolares com plugs mucosos endoluminais.

C) Pneumonia em organização com pneumatoceles.

D) Obstrução bronquiolar com aprisionamento aéreo.

E) Espessamento dos septos interlobulares com vidro fosco de permeio.

E A alternativa correta é a E) Espessamento dos septos interlobulares com vidro fosco de permeio.

Explicação: A atenuação em mosaico é um achado radiológico observado em uma tomografia computadorizada de alta resolução (TCAR) do tórax. Essa aparência é caracterizada pela presença de áreas de atenuação pulmonar diferentes, resultando em um padrão de mosaico na imagem.

A etiologia mais comum da atenuação em mosaico em uma TCAR de tórax é o espessamento dos septos interlobulares com vidro fosco de permeio. Essa alteração indica a presença de inflamação e edema dos espaços intersticiais pulmonares, levando a uma atenuação diferenciada das áreas afetadas.

As outras alternativas apresentam causas diferentes que não estão diretamente relacionadas à atenuação em mosaico:

A) Preenchimento alveolar parcial com espessamento intersticial é mais sugestivo de pneumonia, edema pulmonar ou doenças pulmonares intersticiais difusas.

B) Espessamento de paredes bronquiolares com plugs mucosos endoluminais é mais sugestivo de bronquiectasias ou doença bronquiolite obliterante.

C) Pneumonia em organização com pneumatoceles é uma condição caracterizada por pneumonia com formação de cavidades aéreas.

D) Obstrução bronquiolar com aprisionamento aéreo é mais sugestiva de bronquiolite obliterante.

Portanto, a resposta correta é a alternativa E) Espessamento dos septos interlobulares com vidro fosco de permeio.

Figura 1. Exemplo de uma das respostas corretas do ChatGPT a uma pergunta que avalia habilidades cognitivas de ordem inferior (neste caso, uma definição).

Entre as perguntas que avaliam habilidades cognitivas de ordem superior, o desempenho do ChatGPT foi pior em questões de anatomia, cálculo e classificação e associação entre doenças, porém sem diferença estatisticamente significativa ($p > 0,05$) em comparação com as questões em que se saiu melhor (descrição de achados, manejo clínico e aplicação de conceitos). A Figura 3 mostra um exemplo de erro de uma questão a respeito de anatomia, a Figura 4 mostra um exemplo de acerto em questão que envolve descrição de achados e a Figura 5 mostra um exemplo de acerto em questão que envolve manejo clínico.

Desempenho por tema da questão

O melhor desempenho do ChatGPT foi obtido nas questões de física, sendo estatisticamente superior ao desempenho obtido nas questões clínicas ($p = 0,02$). Dentre as áreas clínicas, o maior índice de acertos foi observado nas questões de pediatria, abdome e tórax, porém sem diferença estatisticamente significativa em comparação com as áreas de pior índice de acertos (ginecologia e obstetrícia e ultrassonografia) ($p > 0,05$).

C Em relação à neoplasia de laringe, assinale a alternativa CORRETA.


A) O carcinoma epidermoide supraglótico costuma ter apresentação clínica mais tardia e maior comprometimento linfonodal que o carcinoma epidermoide glótico.

B) A laringe é o local mais comum de ocorrência do carcinoma epidermoide relacionado ao HPV.

C) A invasão de espaços profundos se dá para os espaços pré-epiglótico, paralaríngeo e parafaríngeo.

D) A invasão da cartilagem cricoide pela neoplasia laríngea a classifica como T3 no TNM.

E) Os linfonodos mais acometidos na neoplasia laríngea são os dos níveis I, II e III.

 A alternativa correta é:

A) O carcinoma epidermoide supraglótico costuma ter apresentação clínica mais tardia e maior comprometimento linfonodal que o carcinoma epidermoide glótico.

Na neoplasia de laringe, o carcinoma epidermoide supraglótico geralmente apresenta sintomas mais tardios em comparação com o carcinoma epidermoide glótico. Isso ocorre porque as lesões supraglóticas podem passar despercebidas inicialmente, já que os sintomas como rouquidão e disfagia são menos perceptíveis. Além disso, o carcinoma epidermoide supraglótico tem maior tendência a apresentar disseminação linfonodal em comparação com o carcinoma epidermoide glótico.

As demais alternativas são incorretas:

B) O carcinoma epidermoide relacionado ao HPV é mais comum em outras regiões, como orofaringe, e não é o mais comum na laringe.

C) A invasão de espaços profundos na neoplasia de laringe não se dá especificamente para os espaços mencionados nessa alternativa.

D) A classificação T3 no TNM se refere a invasão da laringe além da cartilagem tireoide ou invasão da cartilagem cricoides. A invasão isolada da cartilagem cricoide é classificada como T2.

E) Os linfonodos mais acometidos na neoplasia laríngea são os linfonodos cervicais, que são agrupados em diferentes níveis. Os níveis I, II e III correspondem aos linfonodos do pescoço.

Figura 2. Exemplo de uma das respostas corretas do ChatGPT a uma pergunta que avalia habilidades cognitivas de ordem superior (neste caso, a correlação entre doenças).

Desempenho por ano da questão

O melhor desempenho do ChatGPT foi obtido nas questões respondidas por todos os residentes, desde o primeiro ano (61,9%; 57/92), seguido pelas questões respondidas pelos residentes de segundo e terceiro anos (50%; 17/34) e pelas questões respondidas apenas pelos residentes de terceiro ano (36,9%; 14/39), porém não houve diferença estatisticamente significativa ($p > 0.05$).

C Quais são os nervos cranianos que passam pela fissura orbitária superior?


A) I, II, III e IV.

B) II, III, IV e V1.

C) III, IV, V1 e VI.

D) IV, V1, V2 e VI.

E) IV, V1, V2 e V3.

 Os nervos cranianos que passam pela fissura orbitária superior são:

D) IV, V1, V2 e VI.

A fissura orbitária superior é uma abertura localizada na porção superior da órbita ocular. Através dessa fissura, passam os seguintes nervos cranianos:

- Nervo troclear (IV): É responsável pelo movimento do músculo oblíquo superior do olho.
- Ramo oftálmico do nervo trigêmeo (V1): É responsável pela sensibilidade da região da testa, pálpebra superior, parte anterior do couro cabeludo e da córnea.
- Ramo maxilar do nervo trigêmeo (V2): É responsável pela sensibilidade da região da bochecha, lábio superior, parte do nariz, maxila superior e dentes superiores.
- Nervo abducente (VI): É responsável pelo movimento do músculo reto lateral do olho.

Portanto, a opção correta é a D) IV, V1, V2 e VI.

Figura 3. Exemplo de erro do ChatGPT em uma questão de anatomia em neurorradiologia. A resposta correta seria a alternativa C.

Avaliação qualitativa das respostas

A avaliação unânime dos avaliadores foi que o desempenho do ChatGPT foi satisfatório, especialmente considerando-se que seu banco de dados não foi desenvolvido especificamente para uso em radiologia. O alto grau de assertividade da plataforma em fornecer suas respostas, nunca utilizando palavras que indicassem uma possível dúvida ou hesitação (Figuras 1 a 5) também chama a atenção, mesmo em questões em que sua resposta foi a incorreta (Figura 3). Outro fato interessante é que na maior parte das questões (65%; 107/165), não apenas o programa indicou a resposta correta, mas também analisou todas as assertivas, indicando porque as julgou incorretas (Figuras 1, 2 e 4).

DISCUSSÃO

Neste estudo, o primeiro realizado com dados exclusivamente brasileiros, evidencia-se que o ChatGPT ainda não apresenta índice de acertos elevado o suficiente em questões de radiologia para obter a nota exigida para a aprovação na avaliação anual dos residentes em radiologia e diagnóstico por imagem do CBR. O desempenho nas

C Considere um recém-nascido, com 48 horas de vida, com distensão abdominal, não eliminação de mecônio. No RX simples do abdome, observa-se distensão de vários segmentos de alças intestinais, sem níveis hidroaéreos, ausência de ar no reto e efeito de massa no quadrante inferior direito com densidade de partes moles com pequenas áreas de ar entremeadas. Realizado enema opaco, que identificou um cólon medindo 1,0 cm de calibre médio, com pequenas falhas de enchimento. Nesse caso, qual é o provável diagnóstico?

A) Síndrome da rolha meconial.
 B) Íleo meconial.
 C) Volvo de intestino médio.
 D) Banda de Ladd.
 E) Atresia jejunal.

ChatGPT: No caso apresentado, o provável diagnóstico é a letra B) Íleo meconial.

Explicação:

A síndrome da rolha meconial (opção A) geralmente apresenta obstrução distal, com acúmulo de mecônio na ampola retal, o que não foi descrito no caso.

O volvulo de intestino médio (opção C) geralmente ocorre em crianças mais velhas e pode causar obstrução intestinal completa, o que não foi mencionado no caso.

A banda de Ladd (opção D) é uma anomalia congênita relacionada à má rotação do intestino durante o desenvolvimento fetal e pode levar a sintomas de obstrução, mas não é a causa mais provável no caso apresentado.

A atresia jejunal (opção E) também é uma possível causa de obstrução intestinal em recém-nascidos, mas os achados descritos no caso não são típicos dessa condição.

O íleo meconial (opção B) é uma complicação associada à fibrose cística, em que o mecônio é anormalmente espesso e obstrui as alças intestinais. A presença de distensão abdominal, ausência de eliminação de mecônio, distensão de segmentos de alças intestinais no raio-x, cólon estreitado no enema opaco e densidade de partes moles com pequenas áreas de ar entremeadas são características consistentes com o íleo meconial.

Portanto, com base nas informações fornecidas, o provável diagnóstico é íleo meconial. No entanto, é importante ressaltar que o diagnóstico definitivo e o tratamento devem ser realizados por um médico especialista após uma avaliação clínica completa e exames complementares adequados.

Figura 4. Exemplo de acerto do ChatGPT em uma questão em que há descrição dos achados de imagem e solicita-se o diagnóstico.

questões nacionais foi pior do que o observado em questões norte-americanas da mesma especialidade (53,3% naquelas e 69% nestas)⁽⁷⁾, o que pode estar relacionado a diferenças entre as provas de acordo com o conhecimento

C Analise o caso clínico a seguir. Um paciente, sexo masculino, 20 anos de idade, refere uso de esteroides de forma rotineira para aumento da massa muscular e queixa-se de aumento do volume mamário bilateral. A mamografia apresenta apenas densidades fibroglandulares em ambas as regiões retroareolares e a ultrassonografia revelou um nódulo sólido, hipocogênico, heterogêneo, palpável, com 0,6 cm na região subareolar da mama direita. Considerando essas informações, assinale a alternativa correta.

A) Trata-se de um caso de ginecomastia e o paciente deve ser orientado a cessar o uso de esteroides.
 B) Deve ser recomendado o controle em 6 meses do nódulo e reavaliação com mamografia da ginecomastia após redução do volume mamário com a interrupção do uso de esteroides.
 C) Deve ser recomendado o controle evolutivo em 6 meses para o nódulo da mama direita, não sendo necessária a realização de nova mamografia.
 D) Deve ser indicada a biópsia do nódulo da mama direita.
 E) A ressonância magnética é o próximo exame a ser realizado.

ChatGPT: Considerando as informações apresentadas, a alternativa correta é a A) Trata-se de um caso de ginecomastia, e o paciente deve ser orientado a cessar o uso de esteroides.

Explicação: O paciente é um jovem de 20 anos que relata o uso de esteroides para aumento da massa muscular e apresenta aumento do volume mamário bilateral. A mamografia mostra apenas densidades fibroglandulares nas regiões retroareolares, e a ultrassonografia revela um nódulo sólido, hipocogênico, heterogêneo, palpável, com 0,6 cm na região subareolar da mama direita.

Esses achados são compatíveis com ginecomastia, que é o aumento do tecido mamário em homens devido a um desequilíbrio hormonal. O uso de esteroides anabolizantes é uma das causas conhecidas da ginecomastia em homens jovens que praticam musculação. A presença de um nódulo sólido na região subareolar é comum na ginecomastia, e a mamografia geralmente mostra apenas densidades fibroglandulares, sem achados suspeitos.

Portanto, nesse caso, a conduta adequada seria orientar o paciente a cessar o uso de esteroides, uma vez que eles podem estar contribuindo para o desenvolvimento da ginecomastia. Não é necessário realizar uma biópsia do nódulo, pois as características descritas são consistentes com a ginecomastia e não sugerem malignidade. O acompanhamento evolutivo pode ser recomendado para avaliar a regressão da ginecomastia após a interrupção do uso de esteroides, mas não é necessária a realização de nova mamografia. A ressonância magnética não é indicada rotineiramente no diagnóstico da ginecomastia.

Figura 5. Exemplo de acerto do ChatGPT em uma questão em que há descrição de um caso clínico com exame de imagem e solicita-se a conduta mais adequada.

específico que cada país exige dos futuros radiologistas. Novos estudos similares realizados em outros países poderão ajudar a compreender tais diferenças.

A análise das 77 questões que o ChatGPT não acertou mostra que seus erros podem ser atribuídos basicamente ao desconhecimento do assunto que está sendo tratado, como visto na Figura 3. Não foram identificados erros de interpretação do enunciado, associações ilógicas ou ocorrência de alucinações. Esse resultado está de acordo com o que é descrito na literatura, uma vez que alucinações não são tão frequentes em *chatbots* porque estes são elaborados para responder a perguntas com base em regras estabelecidas durante a fase de programação e nas informações contidas em sua base de dados, e não para gerar novas informações⁽¹⁵⁾, o que costuma ser a fonte de alucinações. Estudo semelhante recente também confirmou essa tendência⁽¹⁶⁾, o que sugere que é a falta de familiaridade do

chatbot com as especificidades e nuances da radiologia o principal entrave para um maior índice de acertos.

O melhor desempenho obtido entre as questões que avaliam habilidades cognitivas de ordem inferior sobre as que avaliam habilidades de ordem superior já foi demonstrado na literatura⁽⁷⁾, tendo sido reforçado neste estudo. Este achado mostra a capacidade de a IA em reconhecer e expressar conceitos e definições, mas indica que ainda há avanços a serem obtidos em termos de resolução de desafios mais complexos. É importante que esta característica dos atuais modelos de IA seja conhecida, a fim de que futuros esforços sejam direcionados para elevar seu desempenho em ambos os tipos de ordem cognitiva.

Modelos de linguagem grande como ChatGPT são treinados, a partir de uma ampla base de dados, para reconhecer padrões e relações entre palavras. Dessa forma, a superioridade do índice de acertos em questões de física sobre as questões clínicas observada no presente estudo é compreensível. Uma vez que a base de dados não foi formada para atender especificamente as necessidades do médico radiologista, outras áreas do conhecimento que transcendem essa especialidade, como é o caso da física, tem o potencial de gerar maior número de associações, aumentando seu índice de acertos frente aos desafios propostos. Tais modelos, e o próprio ChatGPT, poderão se beneficiar de maior treinamento nessa especialidade médica, porém, até lá, é importante que o radiologista esteja ciente dessa limitação.

Da mesma forma, a ausência de diferença estatisticamente significativa entre as subespecialidades da radiologia pode ser compreendida pela baixa familiaridade do ChatGPT com os termos e jargões que compõem cada subespecialidade. A radiologia e cada uma das suas subáreas têm vernáculo próprio que é utilizado para elaborar relatórios, classificações e diagnósticos. Enquanto a base de dados dos modelos de linguagem larga não estiver treinada especificamente para lidar com esses termos, a IA pode ser induzida a fazer associações incorretas, o que limita o seu índice de acertos. Por exemplo, a palavra “densidade” tem um significado óbvio para o radiologista, mas que pode ser reconhecido pelo ChatGPT como um conceito diverso do pretendido pelo especialista, simplesmente pela falta de treinamento com o termo no contexto específico. Seu treinamento nessa linguagem técnica específica poderá melhorar o índice de acertos da IA não apenas na radiologia como um todo, mas também em suas subáreas.

Outro achado digno de nota constatado neste estudo é o fato de o ChatGPT ter analisado todas as assertivas da maior parte das questões que lhe foram apresentadas. Não ficou claro qual fator o motivou a realizar tal análise em algumas questões e em outras não, pois o fenômeno foi observado em perguntas de todas as especialidades e tipos, independentemente de suas características. Ainda assim, mesmo quando a análise de todas as alternativas não é feita de forma espontânea, é possível pedir ao ChatGPT

na mensagem seguinte que realize tal avaliação, tendo sido essa solicitação atendida satisfatoriamente neste estudo em 100% das vezes. Essa é uma habilidade que pode vir a ser útil para os residentes que desejam utilizar as questões de provas antigas disponibilizadas pelo CBR como um material de estudo. Mais do que simplesmente indicar a alternativa correta, a plataforma tende a fornecer um estudo completo das assertivas que compõe a pergunta, revisando os temas nela abordados, o que indica um possível papel do ChatGPT como ferramenta auxiliar de estudo, capaz de revisar de forma sucinta, mas eficiente, temas de interesse do residente em radiologia.

Uma das diferenças que este trabalho apresentou em comparação com similares realizados no exterior inclui um maior número de respostas corretas em física (90%) do que em estudo realizado nos Estados Unidos (40%)⁽⁷⁾. Embora não se possa afirmar com certeza, pode-se questionar se diferenças no conteúdo das perguntas (variações nos tópicos dentro do campo da física que são cobrados em cada país) ou em seu processo de formulação (neste estudo, elas foram criadas por uma comissão especializada do CBR, uma instituição nacional, e no estudo dos Estados Unidos por pesquisadores de um único centro), poderiam contribuir para essa divergência. Além disso, embora ainda não esteja claro, é possível que o idioma também exerça alguma influência na performance do ChatGPT, uma vez que há maior disponibilidade de literatura em língua inglesa para treinamento do modelo e, teoricamente, menor familiaridade com perguntas em português. Além disso, a tradução realizada pela máquina pode não capturar perfeitamente o sentido de alguns dos termos ou expressões naturais em língua portuguesa. Conforme novas publicações em diferentes línguas forem surgindo, espera-se que esse tópico seja elucidado.

Este nosso estudo apresenta limitações. Foram utilizadas apenas questões objetivas teóricas que não envolviam a interpretação de imagens radiológicas, pois o ChatGPT ainda não possui essa capacidade. É possível que o fato de ter sido corrigido após seus erros tenha exercido influência sobre o desempenho do ChatGPT, uma vez que talvez suas respostas tivessem sido diferentes caso não houvesse tal retorno. O quanto essa interação com a máquina impacta o resultado final ainda é um campo de estudo a ser explorado. O número de questões em cada subárea foi relativamente pequeno, o que limita a comparação entre esses grupos. Futuros estudos com maior número de questões poderão enriquecer essa discussão.

CONCLUSÃO

Este estudo mostra que, mesmo sem treinamento dedicado a essa área, o ChatGPT apresenta desempenho razoável, mas ainda insuficiente para aprovação, em questões de radiologia formuladas pelo CBR. Espera-se que o treinamento específico em radiologia de modelos de IA como o ChatGPT venha a tornar seu desempenho em

questões dessa especialidade progressivamente melhor, devendo a comunidade radiológica permanecer atenta a essa evolução para aproveitar o seu potencial.

Agradecimento

Os autores agradecem a Comissão de Admissão e Titulação do CBR, na pessoa do Dr. Tulio Augusto Alves Macedo, pela autorização do uso das questões formuladas pelo CBR neste estudo.

REFERÊNCIAS

1. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. *Yearb Med Inform.* 2019;28:16–26.
2. Morreel S, Mathysen D, Verhoeven V, Aye, AI! ChatGPT passes multiple-choice family medicine exam. *Med Teach.* 2023;45:665–6.
3. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof.* 2023;20:1.
4. Rao A, Kim J, Kamineni M, et al. Evaluating ChatGPT as an adjunct for radiologic decision-making. [Preprint]. *medRxiv.* 2023:2023.02.02.23285399.
5. G1 Globo.com. O que é ChatGPT e por que alguns o veem como ameaça? [cited 2023 June 10]. Available from: <https://g1.globo.com/tecnologia/noticia/2023/01/19/o-que-e-chatgpt-e-por-que-alguns-o-veem-como-ameaca.ghtml>.
6. OpenAI. Introducing ChatGPT. [cited 2023 June 2]. Available from: <https://openai.com/blog/chatgpt/>.
7. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology.* 2023;307:e230582.
8. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* 2023;6:1169595.
9. Biswas S. ChatGPT and the future of medical writing. *Radiology.* 2023;307:e223312.
10. Gilson A, Safranek CW, Huang T, 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;9:45312.
11. Weng TL, Wang YM, Chang S, et al. ChatGPT failed Taiwan's Family Medicine Board Exam. *J Chin Med Assoc.* 2023;86:762–6.
12. Jeblick K, Schachtner B, Dext J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *arXiv 2212.14882 [cs.CL].* [cited 2023 June 2]. Available from: <https://arxiv.org/abs/2212.14882>.
13. Colégio Brasileiro de Radiologia e Diagnóstico por Imagem. Avaliação anual de residentes – provas anteriores. [cited 2023 Jun 2]. Available from: <https://cbr.org.br/avaliacao-anual-de-residentes-provas-anteriores/>.
14. Qasrawi R, BeniAbdelrahman A. The higher and lower-order thinking skills (HOTS and LOTS) in unlock English textbooks (1st and 2nd editions) based on Bloom's taxonomy: an analysis study. *International Online Journal of Education and Teaching.* 2020;7:744–58.
15. Alkaiisi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15:e35179.
16. Patil NS, Huang RS, van der Pol CB, et al. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Can Assoc Radiol J.* 2023;8465371231193716.

