UNIVERSIDADE FEDERAL DO PARANÁ

CLEVERSON ALEX LEITÃO

# DESEMPENHO DO CHATGPT EM QUESTÕES DE RADIOLOGIA E DE CONCURSOS DE RESIDÊNCIA MÉDICA

CURITIBA
2024

CLEVERSON ALEX LEITÃO

# DESEMPENHO DO CHATGPT EM QUESTÕES DE RADIOLOGIA E DE CONCURSOS DE RESIDÊNCIA MÉDICA

Tese apresentada ao curso de Pós-Graduação em Medicina Interna, Setor de Ciências da Saúde, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Medicina Interna.

Orientador: Prof. Dr. Dante Luiz Escuissato
Coorientadora: Profa. Dra. Lêda Maria Rabelo

CURITIBA
2024

L533 Leitão, Cleverson Alex
Desempenho do chatgpt em questões de radiologia e de concursos de residência médica [recurso eletrônico] / Cleverson Alex Leitão. – Curitiba, 2024.

Tese (doutorado) – Programa de Pós-Graduação em Medicina Interna e Ciências da Saúde. Setor de Ciências da Saúde. Universidade Federal do Paraná.
Orientador: Prof. Dr. Dante Luiz Escuissato
Coorientadora: Profa. Dra. Lêda Maria Rabelo

1. Inteligência artificial. 2. Educação médica. 3. Radiologia. 4. Questões de prova. I. Escuissato, Dante Luiz. II. Rabelo, Lêda Maria. III. Programa de Pós-Graduação em Medicina Interna e Ciências da Saúde. Setor de Ciências da Saúde. Universidade Federal do Paraná. IV. Título.

FICHA CATALOGRÁFICA ELABORADA PELO SISTEMA DE BIBLIOTECAS/UFPR
BIBLIOTECA DE CIÊNCIAS DA SAÚDE, BIBLIOTECÁRIA: RAQUEL PINHEIRO COSTA JORDÃO CRB 9/991

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS DA SAÚDE
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO MEDICINA INTERNA E CIÊNCIAS DA SAÚDE - 40001016012P1

# TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MEDICINA INTERNA E CIÊNCIAS DA SAÚDE da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **CLEVERSON ALEX LEITÃO** intitulada: **"DESEMPENHO DO CHATGPT EM QUESTÕES DE RADIOLOGIA E DE CONCURSOS DE RESIDÊNCIA MÉDICA"**., sob orientação do Prof. Dr. DANTE LUIZ ESCUISSATO, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 29 de Fevereiro de 2024.

Assinatura Eletrônica
02/03/2024 19:12:05.0
DANTE LUIZ ESCUISSATO
Presidente da Banca Examinadora

Assinatura Eletrônica
04/03/2024 19:23:01.0
ROSÂNGELA REQUI JAKUBIAK
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

Assinatura Eletrônica
05/03/2024 16:28:17.0
EDUARDO ANTONIO ANDRADE DOS SANTOS
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

Assinatura Eletrônica
01/03/2024 17:59:29.0
BERNARDO CORRÊA DE ALMEIDA TEIXEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
01/03/2024 20:33:20.0
LÊDA MARIA RABELO
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

A Deus, a fonte suprema de sabedoria, e à ciência, a manifestação humana do desejo de compreender Seus mistérios.

## AGRADECIMENTOS

Agradeço a Deus pela vida, pela saúde e pela disposição de seguir em frente mesmo diante dos desafios da vida.

A minha mãe, Divina, pelo estímulo constante ao estudo e ao desenvolvimento pessoal.

A meu pai, Carlos (*in memorian*), por quase todos os traços da minha personalidade.

A minha esposa, Marjorie, a minha melhor metade, sem a qual nada mais importa.

A meu filho, Antônio, a alegria dos meus dias.

A meus irmãos, Cibelle e Cleber, pelo companheirismo e paciência fraternais.

A meu orientador, professor doutor Dante Luiz Escuissato, pelo exemplo de conhecimento, retidão moral e dedicação à carreira acadêmica.

A minha coorientadora, professora Lêda Maria Rabelo, pela maestria como professora e como médica.

Aos colegas da banca, professores doutores Rosângela, Eduardo e Bernardo, por me acompanharem em diferentes momentos ao longo de minha trajetória profissional, me estendendo a mão sempre que precisei.

A todos os outros professores que me acompanharam e incentivaram ao longo da minha formação, especialmente Jéferson, Marcus, doutor Rui Fernando Pilotto e doutora Angélica Boldt.

A meu grande amigo, Gabriel Lucca de Oliveira Salvador, sem o qual esse trabalho não teria sido possível.

A todos os outros amigos, desde a infância até a vida adulta. Impossível citar o nome de todos, mas igualmente impossível deixar de mencionar nomes como Ayslan, Bruno, Matheus, Felipe, Christian, Denilson, Diego, Diego, Raphael, Farouk, Eduardo e Patrícia.

A meus alunos, passados, presentes e futuros, razão de ser deste trabalho e todos os demais trabalhos acadêmicos.

“Amor é o presente dado sem esperança de retorno.” (Santo Antônio de Pádua e Lisboa, doutor da Igreja.)

## RESUMO

Recentes avanços da inteligência artificial levantam questionamentos sobre seu uso na medicina, mas seu potencial para auxiliar na educação médica ainda é incerto. O ChatGPT, chatbot online publicamente disponível, constitui um desses avanços. Compreender seu grau de conhecimento médico pode auxiliar a definir seu potencial uso futuro não só na medicina, mas também em educação médica. Este trabalho objetiva testar a acurácia do ChatGPT na resolução de questões de concursos de residência médica e em questões da avaliação anual dos residentes de radiologia pelo Colégio Brasileiro de Radiologia e analisar, com base nesta performance, suas implicações na formação médica. 500 questões de concursos de residência médica do Hospital de Clínicas da Universidade Federal do Paraná (2017-2021) e 165 questões da avaliação anual dos residentes do CBR (2018, 2019 e 2022) foram apresentadas ao ChatGPT, e suas respostas, analisadas. Elas foram divididas, para análise estatística, em questões que avaliavam habilidades cognitivas de ordem superior ou inferior e de acordo com a especialidade e tipo de questão. Foram realizados testes t de student, chi-quadrado e ANOVA. O ChatGPT acertou 56% das questões de residência (280/500) e 53,3% das questões de radiologia (88/165). Houve diferença estatisticamente significativa comparando-se o desempenho em questões que avaliam habilidades cognitivas de ordem inferior e superior tanto nas questões de residência (63,8% x 52,5%, p= 0,01) quanto nas questões de radiologia (64,4 x 47,2%, 50/106, p= 0,01). Nas questões de residência, melhor performance foi obtida em clínica médica (70%), seguida por pediatria (57%), cirurgia geral (56%), ginecologia e obstetrícia (51%) e medicina preventiva e social (46%), sem diferença significativa entre essas áreas, mas com significância entre os diferentes tipos de perguntas ($p < 0,01$). O desempenho do ChatGPT foi acima da linha de corte de quatro das especialidades dos concursos de residência. Em radiologia, houve maior índice de acertos em física (90%, 18/20) do que em questões clínicas (48%, 70/145) ($p = 0,02$). Não houve diferença significativa de desempenho entre subespecialidades ou ano de residência ($p>0,05$). O conhecimento exibido mostra que é possível a um estudante utilizar o ChatGPT para revisar diferentes temas, especialmente porque ele costuma realizar uma análise aprofundada de todas as assertivas de cada questão. Ele também possui valioso poder de síntese, sendo capaz de sumarizar informações disponíveis em consensos e diretrizes, facilitando a acesso do estudante a essas informações. Em resumo, o ChatGPT apresentou desempenho similar ao de candidatos reais em questões de concursos de residência e em questões que avaliam residentes de radiologia brasileiros. A análise de suas respostas e habilidades sugere um possível papel auxiliar dessa ferramenta para acadêmicos de medicina e médicos residentes estudarem os assuntos abordados.



Palavras-chave: Inteligência artificial; educação médica; questões de prova; radiologia.

## ABSTRACT

Recent advances in artificial intelligence raise questions about its use in medicine, but its potential to assist in medical education is still uncertain. ChatGPT, a publicly available online chatbot, is one such advancement. Understanding its level of medical knowledge can help determine its potential future use not only in medicine but also in medical education. This study aims to test the accuracy of ChatGPT in answering questions from medical residency exams and in questions from the annual evaluation of radiology residents performed by the Brazilian College of Radiology. Based on this performance, the implications for medical education are analyzed. 500 medical residency exam questions from the Clinical Hospital of the Federal University of Paraná (2017-2021) and 165 questions from the annual evaluation of radiology residents by the Brazilian College of Radiology (2018, 2019, and 2022) were presented to ChatGPT, and its responses were analyzed. For statistical analysis, questions were divided based on whether they assessed higher or lower-order cognitive skills and according to specialty and question type. Student's t-tests, chi-square tests, and ANOVA were conducted. ChatGPT correctly answered 56% of residency examination questions (280/500) and 53.3% of radiology questions (88/165). There was a statistically significant difference in performance between questions assessing lower and higher-order cognitive skills in both residency examination questions (63.8% vs. 52.5%, p=0.01) and radiology questions (64.4% vs. 47.2%, 50/106, p=0.01). In residency questions, the best performance was in internal medicine (70%), followed by pediatrics (57%), general surgery (56%), gynecology and obstetrics (51%), and preventive and social medicine (46%), with no significant difference between these areas, but with significance between different question types (p<0.01). ChatGPT's performance exceeded the passing threshold for four residency exam specialties. In radiology, there was a higher accuracy rate in physics questions (90%, 18/20) compared to clinical questions (48%, 70/145) (p=0.02). There was no significant difference in performance between subspecialties or residency year (p>0.05). The displayed knowledge indicates that students can use ChatGPT to review various topics, especially because it tends to conduct a thorough analysis of all answer choices for each question. It also has valuable summarization capabilities, being able to condense information from consensuses and guidelines, facilitating student access to this data. In summary, ChatGPT demonstrated a performance similar to that of real candidates in residency examination questions and questions evaluating Brazilian radiology residents. The analysis of its responses and skills suggests a potential supportive role for this tool in the study of medical topics by medical students and residents.



Keywords: Artificial intelligence; medical education; examination questions; radiology.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS OU SIGLAS

| | |
|---|---|
| BERT | *Bidirectional Encoder Representations from Transformers* |
| CADe | *Computer Assisted Detection* (Detecção assistida por computador) |
| CBR | Colégio Brasileiro de Radiologia |
| ChatGPT | *Chat Generative Pre-trained Transformer* |
| EUA | Estados Unidos da América |
| IA | Inteligência Artificial |
| LaMDA | Modelo de linguagem para aplicação em diálogo |
| NLG | *Natural Language Generation* (Geração de linguagem natural) |
| NLP | *Natural Language* Processing (Processamento de linguagem natural) |
| NLU | *Natural Language Transsforming* (Compreensão de linguagem natural) |
| GPT | *Generative Pre-trained Transformer* |
| PBL | *Problem-based* learning (Aprendizado baseado em problemas) |

**SUMÁRIO**

# 1 INTRODUÇÃO

## 1.1 CONTEXTO E PROBLEMA

"Inteligência Artificial" (IA) é o nome geral dado a métodos de computação que simulam o padrão de aprendizado do intelecto humano (WANG; PREININGER, 2019). Os rápidos avanços obtidos recentemente nesse campo do conhecimento têm levado a questionamentos acerca de como ela impactará as mais diversas profissões no futuro, inclusive a medicina. Dentre os produtos de inteligência artificial já existentes, o ChatGPT (*Chat Generative Pre-trained Transformer*) vem ganhando destaque não apenas na literatura científica (MORREEL et al., 2023), mas também na mídia comum (G1, 2023). Trata-se de uma ferramenta de inteligência artificial baseada em relações entre algoritmos de IA com a linguagem humana denominada *Natural Language Processing* (NLP) disponível publicamente desde o dia 30 de novembro de 2022 (OPENAI, 2023). O algoritmo mais recente disponível durante a realização deste estudo era o GPT-3.5, um modelo de linguagem grande treinado com mais de 45 terabytes de dados textuais (atualmente, já há um modelo mais avançado, o GPT-4.0). Por meio de redes neurais, tais dados permitem que a ferramenta seja capaz de analisar e gerar textos similares aos escritos por humanos. Embora não tenha sido treinado especificamente para uso médico, estudos já têm demonstrado seu promissor papel tanto na prática médica (como no auxílio da redação de laudos estruturados em radiologia – ADAMS, 2023), quanto na escrita acadêmica em medicina (BISWAS, 2023; KITAMURA, 2023). 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 (GILSON et al., 2023) e questões para a obtenção de títulos de especialista em radiologia no Canadá/Estados Unidos (BHAYANA; KRISHNA; BLEAKNEY, 2023) e em medicina de família em Taiwan (WENG et al., 2023), com resultados que mostram um desempenho, em geral, próximo ao necessário para aprovação.

Apesar dos vários estudos recentemente realizados sobre o uso do ChatGPT na medicina, há ainda poucos trabalhos a respeito de seu papel na educação médica, e nenhum com dados nacionais. A capacidade da ferramenta de dialogar com um estudante como se fosse um outro aluno, de simular um paciente e de promover

*feedback* para as perguntas recebidas representam potenciais modos de torná-la útil no processo de formação de novos médicos, desde que o conhecimento da plataforma acerca dos temas em estudo seja adequado (GILSON et al., 2023). Este trabalho busca avaliar essas questões na realidade brasileira, por meio da testagem da performance do ChatGPT em temas frequentemente explorados nos concursos para a admissão em um programa de residência médica nacional e em questões para a avaliação da formação de médicos residentes em radiologia, analisando suas respostas, e como elas podem ser utilizadas para auxiliar na formação dos estudantes de medicina e residentes em radiologia.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo geral

Testar a acurácia de uma plataforma de inteligência artificial amplamente disponível (ChatGPT) na resolução de questões aplicadas em exames para admissão em um programa de residência médica e em questões para a avaliação da formação de médicos residentes em radiologia, analisando se há potencial dessa ferramenta em ser utilizada como auxiliar da educação médica.

### 1.2.2 Objetivos específicos

Comparar o desempenho do ChatGPT nas diferentes grandes áreas da medicina (clínica médica, cirurgia geral, ginecologia/obstetrícia, pediatria e medicina preventiva/social) e da radiologia (abdome, tórax, mama, neurorradiologia, pediatria, musculoesquelético, meios de contraste, ultrassonografia, ginecologia e obstetrícia e miscelânea), a fim de verificar se há diferença significativa em seu desempenho a favor de alguma especialidade.

Avaliar sua performance em diferentes modelos de questões, analisando também a diferença entre perguntas que avaliam habilidades cognitivas de ordem superior *versus* inferior, para averiguar se há inclinação favorável a algum tipo de questão.

Analisar a utilidade das respostas fornecidas pelo ChatGPT como ferramenta de estudo para acadêmicos de medicina, médicos se preparando para concursos de residência e residentes de radiologia.

## 1.3 JUSTIFICATIVA

O uso da inteligência artificial nos mais diversos campos do conhecimento humano tem se tornado mais frequente a cada dia, portanto, faz-se necessário compreender desde já seu potencial como auxiliar tanto no ofício médico quanto na formação de novos médicos.

# 2 REVISÃO DA LITERATURA

## 2.1. INTELIGÊNCIA ARTIFICIAL

A história da inteligência artificial (IA) é marcada por tentativas, avanços e desafios que remontam ao século XX. A busca por criar máquinas capazes de realizar tarefas cognitivas humanas começou com os primeiros computadores, mas o termo "inteligência artificial" foi cunhado somente em 1956, por John McCarthy, durante uma conferência em Dartmouth College (MCCARTHY, 1955; MOOR, 2006). Os pioneiros da IA, incluindo nomes como Alan Turing, John McCarthy e Marvin Minsky, buscavam desenvolver algoritmos e sistemas capazes de simular o pensamento humano (TURING, 1950).

Nas décadas seguintes, houve avanços significativos em áreas como o processamento de linguagem natural, a resolução de problemas complexos e o aprendizado de máquina. Nos anos 60, os pesquisadores concentraram-se na IA simbólica, que envolvia o uso de símbolos para representar conhecimento e solução de problemas (GARNELO; SHANAHAN, 2019). Programas de IA iniciais foram projetados para manipular símbolos e regras, levando ao desenvolvimento de sistemas especialistas que podiam imitar as habilidades de tomada de decisão de especialistas humanos em domínios específicos.

Apesar do entusiasmo inicial, a década de 1970 viu uma queda na pesquisa de IA devido a metas ambiciosas e expectativas não atendidas (KAUL; ENSLIN; GROSS, 2020). Ainda assim, os pesquisadores continuaram a trabalhar utilizando bancos de dados de conhecimento especializado para treinar as máquinas, a fim de tentar conseguir fazê-las realizar inferências e resolver problemas.

A década de 1980 testemunhou um ressurgimento do interesse em IA, com o surgimento da pesquisa em redes neurais. Redes neurais, inspiradas na estrutura do cérebro humano, tornaram-se uma abordagem popular para aprendizado de máquina

(DREYFUS, 1990). Tal avanço na capacidade computacional e a disponibilidade de grandes conjuntos de dados contribuíram para a adoção generalizada de aplicações de IA na década de 1990. Tecnologias de NLP (*Natural Language Processing* - Processamento de Linguagem Natural) e visão computacional viram progresso notável, levando ao desenvolvimento de sistemas capazes de entender e interpretar linguagem humana e informações visuais (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Os anos 2000 marcaram a ascensão do aprendizado de máquina *(machine learning)*, de forma que algoritmos podiam aprender com dados a eles fornecidos e aprimorar seu desempenho ao longo do tempo (KOOHY, 2012). Máquinas de vetores de suporte, árvores de decisão e, posteriormente, técnicas de aprendizado profundo ganharam destaque. Grandes conjuntos de dados e o aumento da capacidade computacional impulsionaram ainda mais o crescimento das aplicações de IA. A década de 2010 testemunhou uma revolução na IA com a adoção generalizada do aprendizado profundo, uma subárea do aprendizado de máquina que utiliza redes neurais com múltiplas camadas (KIM et al., 2019). Avanços em processamento de linguagem natural, reconhecimento de imagem e reconhecimento de fala, impulsionados pelo aprendizado profundo, levaram ao desenvolvimento de sistemas avançados de IA, como o *Watson* da IBM e o *AlphaGo* da Google.

Hoje, a IA é uma parte integral da vida diária da sociedade, influenciando setores como saúde, finanças, transporte e entretenimento. A integração da IA em robótica, veículos autônomos e dispositivos inteligentes continua a moldar o futuro. Considerações éticas, desenvolvimento responsável de IA e pesquisas contínuas em IA explicável estão agora na vanguarda, enquanto a sociedade testemunha cada vez mais a evolução da inteligência artificial.

## 2.2. MODELOS DE LINGUAGEM GRANDE

Os modelos de linguagem grandes referem-se a uma categoria de algoritmos de inteligência artificial projetados para compreender e gerar texto em linguagem natural, que são capazes de versar a respeito de um assunto questionado mesmo sem terem recebido treinamento específico para ele (THIRUNAVUKARASU et al., 2023). Esses modelos são treinados com grandes conjuntos de dados para aprender

padrões complexos na linguagem e, assim, são capazes de realizar uma variedade de tarefas relacionadas ao processamento de linguagem natural.

Um exemplo notável de modelos de linguagem grandes é a arquitetura GPT (*Generative Pre-trained Transformer*), desenvolvida pela OpenAI (OPEANAI, 2022). O GPT utiliza uma arquitetura de transformador, que é uma classe de modelos de aprendizado de máquina baseados em atenção. Este e outros modelos semelhantes, como o BERT (*Bidirectional Encoder Representations from Transformers*), têm a capacidade de entender contextos amplos e captar relações semânticas complexas entre palavras e frases (DEVLIN et al., 2018). Isso permite que esses modelos se destaquem em tarefas como tradução automática, resumo de texto, resposta a perguntas, geração de texto e até mesmo em aplicações mais avançadas, como *chatbots* e assistentes de voz.

A característica "grande" nesses modelos geralmente se refere ao elevado número de parâmetros que eles possuem. Modelos como o GPT-3 (terceira iteração do GPT) têm centenas de bilhões de parâmetros, tornando-os extremamente poderosos em termos de capacidade de aprendizado e geração de texto (OPENAI, 2023). Essa escala massiva permite que esses modelos capturem nuances complexas e padrões sutis na linguagem.

No entanto, é importante observar que, embora esses modelos sejam impressionantes em muitas tarefas, eles também enfrentam desafios, como o potencial de gerar respostas tendenciosas e a necessidade de enormes recursos computacionais para treinamento e inferência (TAMKIN et al., 2021). Além disso, questões éticas relacionadas ao uso responsável desses modelos também estão sendo cada vez mais discutidas pela comunidade de pesquisa em inteligência artificial.

## 2.3. CHATGPT

O ChatGPT é um modelo de linguagem grande desenvolvido pela OpenAI, baseado na arquitetura GPT (*Generative Pre-trained Transformer*). O termo "pré-treinado" indica que o modelo é treinado em uma grande quantidade de dados antes de ser ajustado para tarefas específicas. Ele é parte da família de modelos GPT elaborados pela mesma empresa, que inclui GPT-2, GPT-3, e assim por diante.

A OpenAI foi fundada inicialmente em 2015 por Sam Altman, Elon Musk, Ilya Sutskever e Greg Brockman como uma organização sem fins lucrativos com o objetivo declarado de "avançar a inteligência digital da maneira mais provável de beneficiar a humanidade como um todo" (OPENAI, 2023). A empresa reuniu uma equipe dos melhores pesquisadores no campo da inteligência artificial para buscar a construção da IA geral de forma segura.

Os primeiros anos da OpenAI foram marcados por experimentação rápida (PALIHAPITIYA, 2023). A empresa fez progressos significativos na pesquisa em aprendizado profundo e aprendizado por reforço, e lançou o 'OpenAI Gym' em 2016, um conjunto de ferramentas para desenvolver e comparar algoritmos de aprendizado por reforço (OPENAI, 2016).

A OpenAI demonstrou as capacidades desses algoritmos de aprendizado por reforço por meio de seu projeto 'OpenAI Five' em 2018, que treinou cinco agentes de IA independentes para jogar um jogo complexo *online* de múltiplos jogadores chamado 'Dota 2'. Apesar de operarem de forma independente, esses agentes aprenderam a trabalhar como uma equipe coesa para coordenar estratégias dentro do jogo (OPENAI, 2018).

Um salto de desenvolvimento ocorreu em junho de 2018. A empresa lançou um artigo intitulado "Melhorando a Compreensão da Linguagem por Pré-treinamento Generativo" (RADFORD et al., 2018), que introduziu a arquitetura fundamental para o modelo *Generative Pre-trained Transformer* (GPT ou “GPT-1”).

Em 2019, a OpenAI apresentou o GPT-2 como um modelo de linguagem grande com 1,5 bilhão de parâmetros (os pesos ajustáveis do modelo que são aprendidos durante o treinamento) (OPENAI, 2019). Esse modelo ganhou atenção devido às suas impressionantes capacidades de geração de linguagem natural, mas a OpenAI inicialmente reteve o modelo completo devido a preocupações com seu possível uso indevido na geração de notícias falsas e desinformação. Em novembro de 2019, a OpenAI eventualmente lançou o modelo completo para o público.

Em 3 em junho de 2020, foi lançado o GPT-3. O GPT-3 foi um marco significativo, sendo um dos maiores modelos de linguagem já criados até aquele momento, com 175 bilhões de parâmetros. O GPT-3 demonstrou bom desempenho em tarefas de linguagem natural, como geração de texto, tradução automática, resolução de problemas matemáticos e até mesmo conversas em linguagem natural.

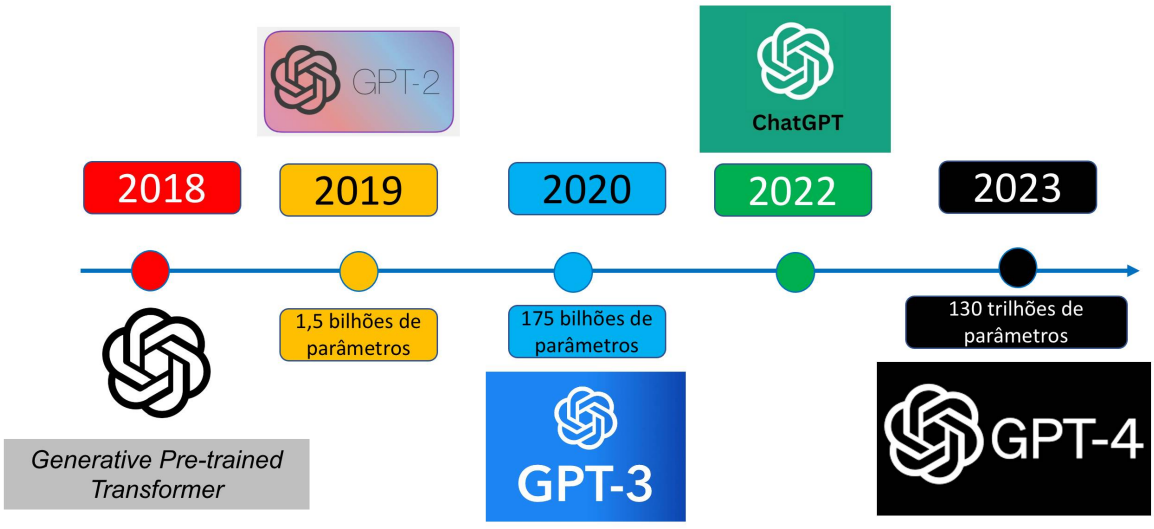
No entanto, o acesso ao GPT-3 inicialmente era limitado, disponível apenas através de liberação pela OpenAI.

Em dezembro de 2020, a OpenAI desenvolveu o ChatGPT como uma versão mais interativa do GPT-3, possibilitando aos usuários experimentar e interagir com o modelo diretamente em seu site. Em novembro de 2022, ele foi modernizado para a versão 3.5 do GPT e seu acesso foi liberado para uso público. Desde então, a OpenAI continuou a aprimorar e iterar sobre os modelos de linguagem, lançando diferentes variantes, cada uma com melhorias em relação às anteriores. Esses modelos são treinados em grandes quantidades de dados textuais da internet, permitindo-lhes gerar respostas contextuais e relevantes em uma variedade de tarefas de linguagem natural.

Desde 2022, Chat-GPT rapidamente capturou a atenção global, tornando-se o aplicativo mais rápido a atingir 100 milhões de usuários em apenas dois meses após o lançamento. Capitalizando esse sucesso, a OpenAI introduziu, em 2023, um modelo de assinatura e revelou seu modelo mais sofisticado até então (com mais de 100 trilhões de parâmetros), o GPT-4 (OPENAI, 2023), exclusivamente para assinantes.

A Figura 1 sintetiza a evolução das diferentes versões do GPT até o modelo mais recente.

FIGURA 1 – EVOLUÇÃO DO GPT.




FONTE: ELABORADA PELOS AUTORES (2024).

O ChatGPT explora diferentes habilidades da inteligência artificial, essencialmente utilizando tanto NLP (*natural language processing* - processamento de linguagem natural) quanto *machine learning* (aprendizado de máquina) (OPENAI, 2023).

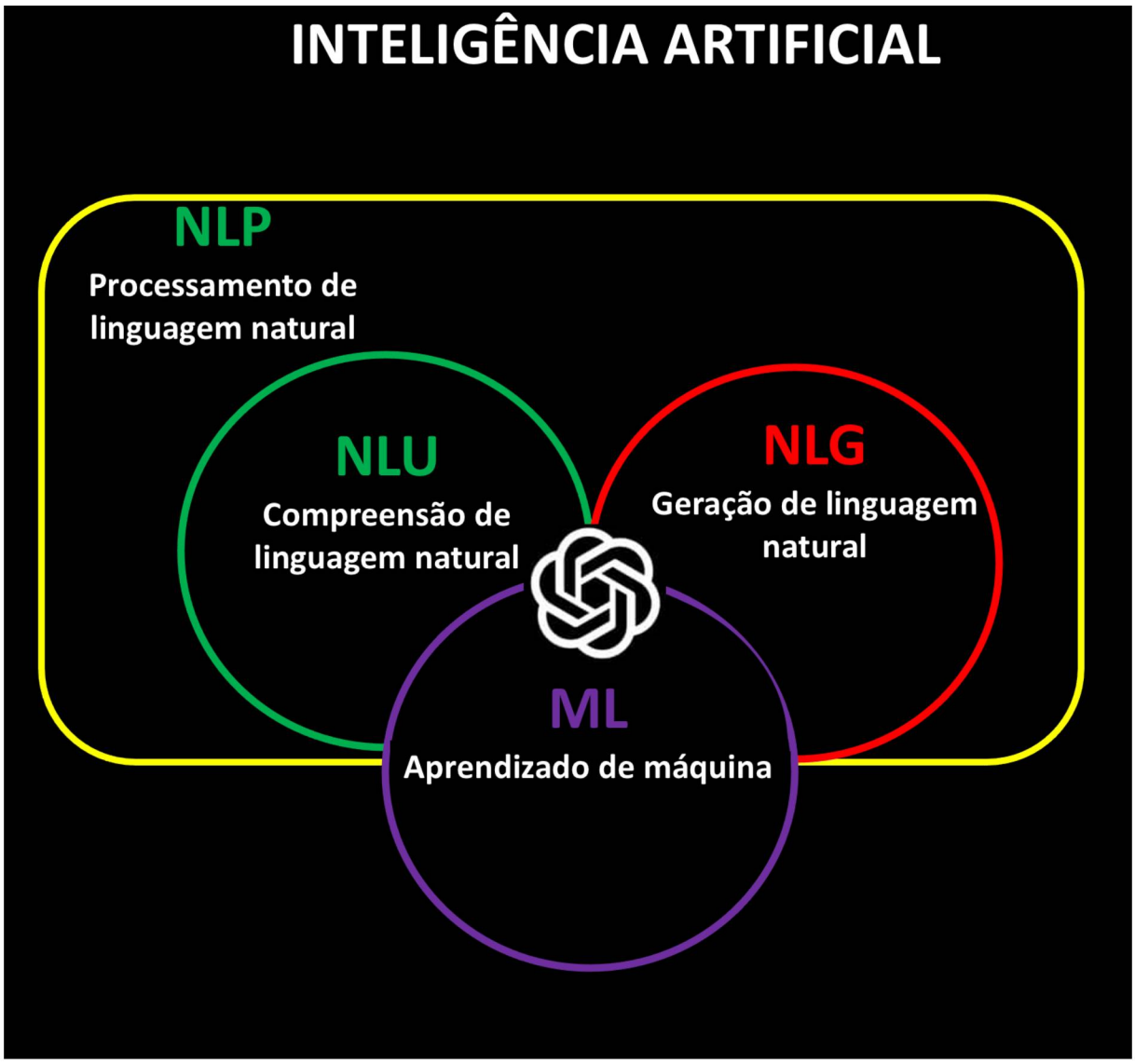
O processamento de linguagem natural é uma área da IA que lida com o processamento e a análise da linguagem humana (LOCKE, 2021). Além da compreensão (NLU – *natural language understanding*), a NLP é capaz de produzir escritas e narrativas a partir de uma base de dados (NLG – *natural language generation*). *Chatbots* interativos, como o ChatGPT, utilizam tanto a NLU quanto a NLG. A NLU é utilizada para entender a estrutura gramatical das sentenças (análise sintática), compreender o significado das palavras e frases em um contexto específico (análise semântica), para reconhecer nomes de pessoas, locais, datas e outros elementos relevantes em uma sentença e para análise de anáforas (entender a referência de pronomes ou outras expressões que retomam um elemento já citado, garantindo que a máquina compreenda corretamente a quem ou a quê se refere uma determinada parte do texto). Já a NLG é usada para transformar informações e conhecimentos previamente adquiridos em expressões linguísticas naturais (respostas a perguntas, produção de texto coerente, explicação de conceitos, criação de textos e tradução de idiomas, por exemplo), permitindo uma comunicação mais eficaz entre máquinas e humanos.

O aprendizado de máquina, por sua vez, é um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem aos sistemas aprenderem padrões a partir de dados sem serem explicitamente programados; ao invés de seguirem regras fixas com as quais foram programados, os modelos de *machine learning* aprendem com exemplos e experiências (RAJKOMAR, 2019), melhorando sua capacidade de realizar tarefas específicas à medida que são expostos a mais dados. Existem diferentes tipos de aprendizado de máquina, incluindo aprendizado supervisionado (o modelo é treinado com um conjunto de dados rotulado e cada exemplo do conjunto de dados possui uma entrada e uma saída desejada), aprendizado não supervisionado (o modelo é treinado com um conjunto de dados não rotulado e o sistema tenta aprender padrões ou estruturas subjacentes nos dados, sem ter saídas específicas para guiar o processo) e aprendizado por reforço (o modelo interage com um ambiente e aprende a realizar ações para maximizar uma recompensa). O modelo de linguagem GPT-3.5, utilizado pelo

ChatGPT, foi treinado usando a abordagem *transfer learning*, uma forma de aprendizado de máquina que expõe o modelo a uma vasta quantidade de textos da internet, o que permite a ele aprender padrões e relações semânticas na linguagem natural. Reconhecendo tais padrões e relações, o ChatGPT consegue gerar respostas contextualmente relevantes com base nas entradas recebidas. Sua funcionalidade é focada na interação linguística e na geração de texto, contudo, não sendo capaz de realizar tarefas fora do escopo de processamento de linguagem natural.

A figura 2 sintetiza os campos da inteligência artificial de que o ChatGPT se utiliza.

FIGURA 2 – CHATGPT E CAMPOS DA INTELIGÊNCIA ARTIFICIAL.




FONTE: ELABORADA PELOS AUTORES (2024).

É importante lembrar que, apesar de o ChatGPT ser capaz de realizar várias tarefas relacionadas ao processamento de linguagem natural, ele apresenta algumas limitações importantes. A primeira, e principal, é a falta de conhecimento atualizado. A base de conhecimento do ChatGPT tem um corte em janeiro de 2022, logo, ele não possui conhecimento sobre eventos ou desenvolvimentos ocorridos após essa data. Além disso, sua capacidade de entender o contexto de uma conversa tem limitações: em diálogos longos ou complexos, ele pode perder o contexto anterior. Há também considerável sensibilidade contextual: suas respostas podem variar dependendo da formulação da pergunta, e, eventualmente, ele pode interpretar um comando de maneira literal ou fora de contexto. Por fim, a própria OpenAI ressalta as limitações do ChatGPT em termos de segurança digital, recomendando que os usuários evitem compartilhar informações pessoais sensíveis, como senhas ou dados financeiros.

Os principais competidores do ChatGPT no mercado atualmente são o Google Bard, desenvolvido especificamente para concorrer com os produtos da OPENAI, e o Chatsonic, que é anunciado como uma ferramenta de IA voltada para produtores de conteúdo digital.

O Bard é um *chatbot* desenvolvido pelo Google, baseado na família de modelos de linguagem grande com enfoque em aplicativos de diálogo (LaMDA - *Language Model for Dialogue Applications*, no termo cunhado pela empresa). Foi criado como uma resposta direta ao ChatGPT da OpenAI, tendo sido lançado em uma versão de testagem em março de 2023 (MANYIKA, 2023). Trata-se um modelo de linguagem factual, o que significa que ele é treinado em um enorme conjunto de dados de texto e código. Isso permite que ele gere texto, traduza idiomas, escreva diferentes tipos de conteúdo criativo e responda a perguntas de forma informativa. A plataforma alega se capaz de gerar diferentes formatos de texto criativo (como poemas, código, roteiros, peças musicais, e-mail e cartas), de traduzir idiomas de forma precisa e natural, de resumir informações complexas de forma clara e concisa, entre outras habilidades.

O Chatsonic é um *chatbot* desenvolvido pela Writesonic (WRITESONIC, 2023), que, diferentemente do ChatGPT, é capaz de utilizar a base de dados do Google para embasar suas respostas (ao passo que o ChatGPT consulta somente sua própria base de dados). Essa plataforma é equipada também com um mecanismo de "persona", que permite ao usuário assumir uma determinada personalidade artificial para responder às suas perguntas (como, por exemplo, *coach* motivacional

ou *personal trainer*). Ele ainda é capaz de reconhecer comandos de voz e gerar áudios e imagens a partir de textos. Seu uso é restrito a assinantes, contudo.

## 2.4. USO DA IA NA MEDICINA

Nos últimos anos, a IA tem sido cada vez mais aplicada na medicina, com potencial para revolucionar a forma como os médicos diagnosticam, tratam e cuidam dos pacientes.

Uma das áreas mais promissoras de aplicação da IA na medicina é o auxílio ao diagnóstico. A IA pode ser usada para analisar grandes quantidades de dados médicos, como exames de imagens, de laboratório, histopatológicos e registros de prontuários, para identificar padrões capazes de levar a um diagnóstico específico. Utilizando o diagnóstico do câncer de pulmão, por exemplo, estudos publicados na revista *Nature* mostram que inteligências artificiais são capazes de analisar tomografias de tórax de baixa dose para rastreamento dessa doença com uma acurácia comparável ou superior a de um radiologista (ARDILA et al., 2019) e também de auxiliar na graduação histológica do adenocarcinoma pulmonar (PAN et al., 2024).

A IA também pode ser usada para avaliar sinais e sintomas de doenças, tentando associá-las a diagnósticos específicos, diferenciando-as de outras patologias. Durante a pandemia de COVID-19, por exemplo, um aplicativo para *smartphone* (*AI4COVID-19*) foi desenvolvido para diferenciar, por meio da análise da tosse de um paciente, essa doença de outras causas de tosse (IRMAN, 2020). Isso pode ser útil para médicos tanto em termos de diagnóstico quanto para rastreamento de uma enfermidade específica.

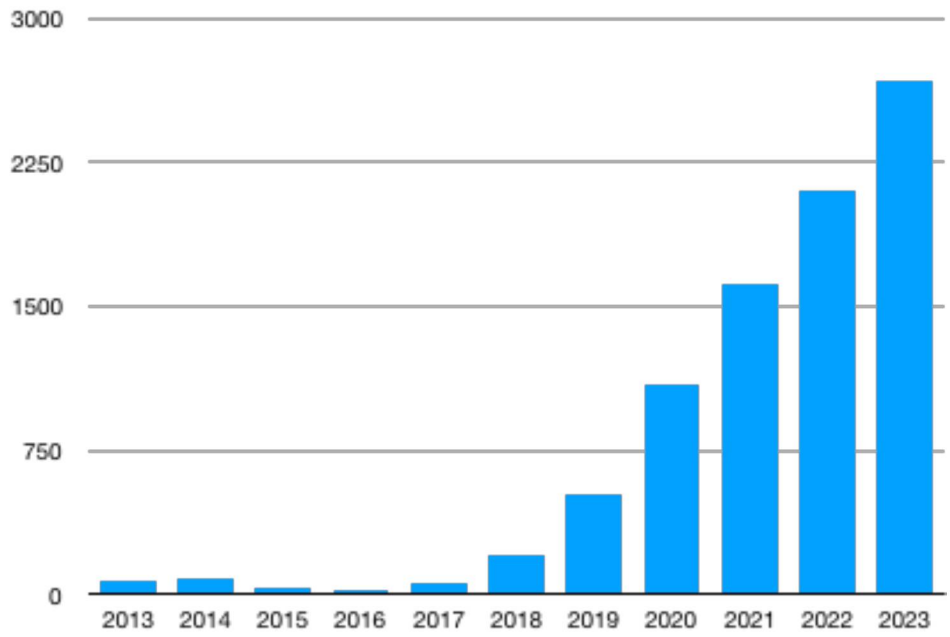
O monitoramento de pacientes com doenças crônicas que necessitam de acompanhamento contínuo, como diabetes, também tem se beneficiado do uso da IA. Modelos mais avançados de sistemas de monitoramento dos níveis séricos de glicose em pacientes com diabetes tipo 1 baseados em inteligência artificial têm sido desenvolvidos (VETTORETTI, 2020; AHMED, 2023), não apenas para permitir um monitoramento em tempo real mais fidedigno da glicemia do paciente, reduzindo as ocorrências de episódios de hipo ou hiperglicemia, mas também para fornecer um plano de insulinoterapia personalizado de acordo com as informações obtidas.

Até mesmo nas áreas que envolvem procedimentos cirúrgicos e diagnósticos invasivos da medicina, a IA tem encontrado campos de atuação. Em oftalmologia, a

IA pode ser utilizada como auxílio não somente no diagnóstico e graduação da catarata, mas também no cálculo do índice de refração esperado no pós-operatório, no acompanhamento em tempo real de imagens geradas durante cirurgias microscópicas e na emissão de alertas importantes para o oftalmologista durante o ato cirúrgico (LINDEGGER; WAWRZYNSKI; SALEH, 2022). A detecção e o diagnóstico de lesões do trato gastrointestinal por endoscopia também podem se beneficiar do uso de redes neurais por meio de sistemas de detecção (CADe – *Computer Assisted detection*) e de diagnóstico assistidos por máquina (CHADEBECQ; LOVAT; STONAYOV, 2023).

Possivelmente a especialidade médica que mais tem se destacado em pesquisas relacionadas ao uso da IA em medicina é a radiologia; centenas de artigos têm sido publicados todos os anos em revistas indexadas versando sobre o papel que ela pode exercer no auxílio da interpretação e reconhecimento de imagens médicas. A figura 3 mostra a evolução no número de artigos encontrados no *pubmed* com os descritores “radiology” e “artificial intelligence”.

FIGURA 3 – EVOLUÇÃO DO NÚMERO DE PUBLICAÇÕES EM IA EM RADIOLOGIA.




FONTE: PUBMED (2024).

## 2.5. IA E EDUCAÇÃO MÉDICA

A educação médica consiste em um esforço contínuo que se estende por toda uma vida, iniciando-se na graduação e estendendo-se não apenas para a pós-graduação e especializações, mas muito além, em um processo que alguns autores denominam de "educação médica continuada" (CHAN; ZARY, 2019). Assim como vem ocorrendo em outras profissões e na própria medicina, o avanço da inteligência artificial nos últimos anos torna o momento auspicioso para a introdução progressiva dessa ferramenta no cotidiano da formação médica.

A maior parte das pesquisas que envolvem o uso da IA em educação médica se baseiam em aprendizado e desenvolvimento de conhecimento, utilizando-se da capacidade da máquina de fornecer *feedback* imediato. O *feedback* é fundamental para que o estudante enxergue suas lacunas de conhecimento e os pontos em que deve se empenhar mais. Contudo, fornecê-lo pode ser desafiador em alguns contextos clínicos inerentes à formação médica, como em um atendimento de emergência. Em um estudo conduzido por Hewson e Little (1998), 80% dos residentes pesquisados relataram nunca terem recebido ou receberem raramente *feedback* corretivo sobre seu desempenho. Um modelo de linguagem grande, por outro lado, pode fornecer *feedback* imediato e formativo sobre o desempenho dos alunos. Em tempos de reforma curricular, em que grande parte das universidades têm substituído o currículo convencional por aprendizado baseado em problemas (*PBL - problem based learning*), a IA pode ter um papel especialmente útil, devido a essa possibilidade de fornecer uma devolutiva passo-a-passo para o estudante frente a um problema.

Uma das aplicações que já têm sido testadas do uso da IA em educação médica é a utilização de um *bot* em substituição a um professor convencional na discussão de casos clínicos e posterior fornecimento de *feedback* a acadêmicos de medicina. McFadden et al. conduziram um estudo em que dois grupos de alunos eram apresentados a diferentes casos de dor articular, tendo o grupo controle recebido uma aula de duas horas e uma sessão de perguntas e respostas com um professor reumatologista e o grupo intervenção recebido uma vídeo-aula resumida e acesso a um *bot* que discutia os mesmos casos clínicos mostrados pelo professor ao grupo controle. O ganho de desempenho no pós-teste em relação ao pré-teste foi superior no grupo com acesso ao *bot* do que no grupo sem acesso (MCFADDEN et al., 2016).

Outro potencial uso para a IA cuja implementação vem sendo testada consiste na sua utilização em avaliações. Dentre as várias formas de avaliar o conhecimento adquirido na faculdade de medicina, as duas mais comuns são os testes escritos e os exames clínicos estruturados (EPSTEIN, 2007). Esses métodos de avaliação geralmente são realizados fora da internet, utilizando papel e caneta, o que limita o uso da IA, pois seria necessário digitalizar uma grande quantidade de material para treinar um modelo satisfatório para esse fim. Com a progressiva adoção de avaliações *online*, é possível que essa barreira seja paulatinamente transposta. Contudo, outros desafios ainda permanecem, como a necessidade de garantir uma via de comunicação segura entre o estudante e a plataforma que aplicará as perguntas e a implementação de medidas capazes de dificultar a cola (SARRAYRIH; ILYAS, 2013). Ademais, o próprio funcionamento da IA precisa ser criteriosamente avaliado, uma vez que erros de programação ou do funcionamento da máquina podem levar a erros na correção das respostas dos estudantes, o que pode acarretar danos graves e indesejados à trajetória acadêmica destes.

Apesar das aplicações que vêm sendo desenvolvidas, ainda há muitos obstáculos para a implementação da inteligência artificial em larga escala nas faculdades de medicina. Entre esses, pode-se citar as dificuldades tecnológicas em adquirir um tamanho de amostra grande o suficiente para o desenvolvimento do modelo de IA que se planeja aplicar; a necessidade de um especialista em conteúdo qualificado e experiente para trabalhar no aprendizado de máquina; e os desafios de comunicação relativos à lacuna de conhecimento entre médicos e programadores (CHAN; ZARY, 2019). O processo para a elaboração de uma IA é multiprofissional, envolvendo, além destes dois profissionais, especialistas em educação e engenheiros de dados, formando um complexo amálgama multifacetário ainda muito distante da realidade de grande parte das universidades brasileiras. Há, ainda, questões éticas que dizem respeito ao sigilo médico e à confidencialidade das informações que são fornecidas ao modelo de IA que não podem ser esquecidas (FRIZE; FRASSON, 2000). Especificamente no campo da ética médica, é esperado que o aprendizado não possa ser completamente automatizado, pois ainda far-se-á necessária a presença de profissionais de saúde capazes de transmitir o conhecimento adquirido pela vivência que a inteligência artificial não é capaz de ter (WARTMAN; COMBS, 2018).

# 3 METODOLOGIA

## 3.1. TIPO DE PESQUISA

Estudo analítico prospectivo realizado entre os dias 31 de maio e 10 de junho de 2023, que não envolveu seres humanos ou dados de pacientes, sendo dispensada a aprovação por comitê de ética institucional.

## 3.2. QUESTÕES UTILIZADAS

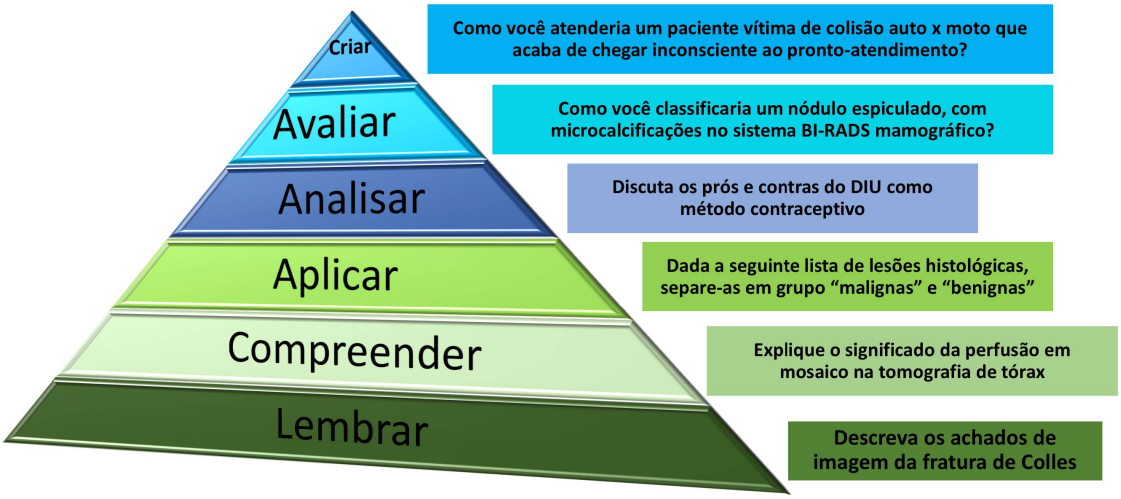
### 3.2.1. Questões de radiologia

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, as quais se encontram disponíveis *online* para acesso público no site do CBR (COLÉGIO BRASILEIRO DE RADIOLOGIA, 2023) 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 a versão 3.5 do ChatGPT 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 subdividas de acordo com os princípios da taxonomia de Bloom. Esses referem-se a uma estrutura hierárquica desenvolvida pelo psicólogo e pedagogo americano Benjamin Bloom e seus colaboradores na década de 1950 (BLOOM, 1956). Ela foi criada com o objetivo de classificar em seis níveis as habilidades cognitivas que os alunos utilizam durante o processo de aprendizado. Os seis níveis são: lembrar (demonstrar a capacidade de lembrar informações, fatos ou conceitos), compreender (explicar ideias ou conceitos usando suas próprias palavras), aplicar (demonstrar a habilidade de usar o conhecimento de uma maneira prática),

analisar (identificar padrões ou tendências), avaliar (julgar, avaliar a validade de uma ideia) e criar (o nível mais alto, em que os alunos demonstram a capacidade de criar algo novo, integrando ideias para solucionar problemas). Eles fornecem uma estrutura útil para pensar sobre o desenvolvimento progressivo das habilidades dos alunos. Neste estudo, a taxonomia foi usada para avaliar o chatGPT de forma similar, dividindo as perguntas em questões que avaliam habilidades cognitivas de ordem inferior (lembrar, aplicar, compreender um conceito) e questões que avaliam habilidades cognitivas de ordem superior (analisar, avaliar e criar algo com o conhecimento obtido). A figura 2 sintetiza e exemplifica tais níveis.

FIGURA 4 – PRINCÍPIOS DA TAXONOMIA DE BLOOM.




LEGENDA: Em tons de azul, as habilidades de ordem cognitiva superior e, em tons de verde, as de ordem inferior.

FONTE: ELABORADA PELOS AUTORES (2024).

Posteriormente, as questões foram novamente divididas de acordo com seu estilo em seis subcategorias: (1) descrição de um achado clínico ou sinal, (2) manejo clínico de um doente, (3) aplicação de um conceito, (4) cálculo ou classificação dos achados descritos, (5) associação entre doenças, (6) 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 (92), questões voltadas para residentes do segundo e terceiro anos (34) e questões voltadas para os residentes do terceiro ano (39).

### 3.2.2. Questões de provas de residência

Foram selecionadas 500 questões dos concursos para a admissão no programa de residência médica do Hospital de Clinicas da Universidade Federal do Paraná realizados entre os anos de 2017 e 2021, as quais se encontram disponíveis *online* para acesso público no site do Núcleo de Concursos dessa instituição (NÚCLEO DE CONCURSOS, 2023). Todas as questões eram do tipo "múltipla escolha", com apenas uma alternativa correta e quatro alternativas falsas. Para garantir a representatividade de todas as grandes áreas da medicina, foram escolhidas 100 questões de cada uma delas: clínica médica, cirurgia geral, pediatria, ginecologia e obstetrícia e medicina preventiva e social.

Posteriormente, essas questões também foram subdividas de acordo com os princípios da taxonomia de Bloom (BLOOM, 1956) em questões que avaliam habilidades cognitivas de ordem inferior e questões que avaliam habilidades cognitivas de ordem superior. Estas foram novamente divididas de acordo com seu estilo em sete subcategorias: (1) descrição de um achado clínico ou sinal, (2) manejo clínico de um doente, (3) aplicação de um conceito, (4) cálculo ou classificação dos achados descritos, (5) revisão fisiopatológica de uma ou mais doenças, (6) indicar o diagnóstico específico e (7) indicar o tratamento específico. Todas as questões foram classificadas independentemente pelos autores do estudo e, nos casos de desavença, uma classificação final foi obtida por consenso.

## 3.3. CHATGPT

Foi utilizada a versão 3.5 do ChatGPT, pois se tratava da mais recente disponível no momento do estudo (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 médicas. O ChatGPT não realiza buscas na internet, respondendo às perguntas utilizando apenas a sua base de dados.

## 3.4. COLETA E ANÁLISE DE DADOS

As questões e suas respectivas alternativas foram apresentadas ao ChatGPT sequencialmente, uma a uma, exatamente 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. Não foi solicitada à máquina que fornecesse uma referência para a resposta escolhida. 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.

## 3.5. 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 descrito acima) e exibida a porcentagem final dessa razão.

A comparação entre o índice de acertos nos diferentes grupos (baixa ordem x alta ordem; física x clínica; tipo de questão) utilizou os testes exato de Fisher e chi-quadrado. A análise entre subgrupos de questões (por tema e ano da questão) utilizou o teste de ANOVA. O programa utilizado foi o STATA 16 e pós-processamento foi realizado no Microsoft Excel 365, com seu pacote de análise de dados. Consideraram-se como estatisticamente significantes valores de $p < 0,05$.

# 4 APRESENTAÇÃO DOS RESULTADOS

## 4.1. RESULTADO GERAL

### 4.1.1. Questões de radiologia

O ChatGPT acertou 53% das questões que lhe foram apresentadas (88 de 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.

TABELA 1 - DESEMPENHO DO CHATGPT DE ACORDO COM O TIPO E O TEMA DA QUESTÃO EM RADIOLOGIA.

| Tipo/tema de questão | Total de questões | Acertos (%) | Valor de p |
|---|---|---|---|
| **Total** | 165 | 88 (53,3) | 0,01* |
| 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) | 0,58* |
| **Tema** | | | 0,02* |
| Física | 20 | 18 (90) | |
| Clínica | 145 | 68 (46,8) | 0,41* |
| Abdome | 20 | 13 (65) | 0,62** |
| Tórax | 15 | 9 (60) | 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) | 0,94** |
| Ultrassonografia | 15 | 3 (20) | 0,78** |
| Pediatria | 15 | 10 (66,7) | 0,93** |
| Gineco-obstetrícia | 10 | 2 (20) | 0,72** |
| Miscelânea | 10 | 4 (40) | 0,65** |

* Teste exato de Fisher;

** Teste ANOVA.

FONTE: ELABORADA PELOS AUTORES (2023).

### 4.1.2. Questões de residência médica

O ChatGPT teve acurácia geral de 56% (acertou 280 de 500 questões apresentadas). Considerando-se a média dos cinco anos, essa nota permitiria a um candidato a residência progredir para a segunda fase do certame em quatro das dezesseis especialidades médicas ofertadas. A TABELA 2 compara a nota obtida pela plataforma com a média nos mesmos cinco concursos da nota de corte obtida pelos candidatos a diferentes especialidades.

TABELA 2 – COMPARAÇÃO ENTRE O DESEMPENHO DO CHATGPT E A NOTA DE CORTE PARA AVANÇO DE FASE NO CONCURSO DE RESIDÊNCIA MÉDICA EM DIFERENTES ESPECIALIDADES MÉDICAS.

| **Especialidade** | **Média da nota de corte (2017 – 2021)** |
|---|---|
| Dermatologia | 71 |
| Oftalmologia | 68,6 |
| Clínica médica | 68 |
| Otorrinolaringologia | 68 |
| Neurocirurgia | 67,3 |
| Anestesiologia | 65,6 |
| Cirurgia Geral | 66,3 |
| Psiquiatria | 64,3 |
| Neurologia | 64 |
| Pediatria | 62 |
| Ginecologia/Obstetrícia | 61 |
| Radiologia | 60,3 |
| **ChatGPT** | **56** |
| Infectologia | 53,6 |
| Ortopedia | 51,3 |
| Patologia | 49 |
| Medicina de Família | 40 |

FONTE: ELABORADA PELOS AUTORES (2023).

A TABELA 3 compara o desempenho do ChatGPT nas questões de concursos de residência e nas questões de radiologia formuladas pelo CBR.

TABELA 3 – COMPARAÇÃO DO DESEMPENHO DO CHATGPT NAS QUESTÕES DE RESIDÊNCIA MÉDICA E NAS QUESTÕES DE RADIOLOGIA

| | **Questões de residência** | **Questões de radiologia** |
|---|---|---|
| **Número de acertos (%)** | 280 (56%) | 88 (53%) |
| **Número de erros (%)** | 220 (44%) | 75 (47%) |

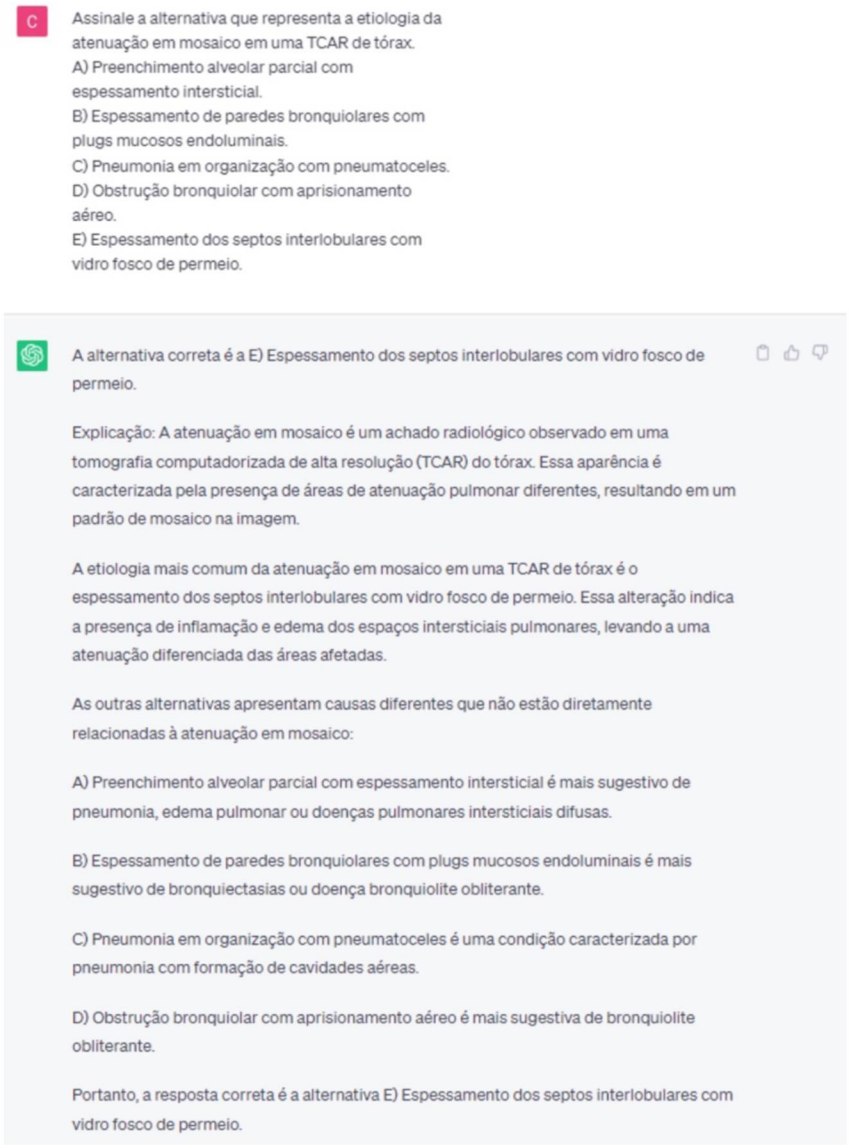
FONTE: ELABORADA PELOS AUTORES (2023).

## 4.2. DESEMPENHO POR TIPO DE QUESTÃO

### 4.2.1. Questões de radiologia

O desempenho em questões que avaliam habilidades cognitivas de ordem inferior (64,4%, 38 de 59) foi superior àquele visto nas questões que avaliam habilidades cognitivas de ordem superior (47,2%, 50 de 106) (p = 0,01). A FIGURA 5 mostra um exemplo de acerto de uma questão que avalia habilidade de ordem inferior e a FIGURA 6 mostra um exemplo de acerto de ordem superior.

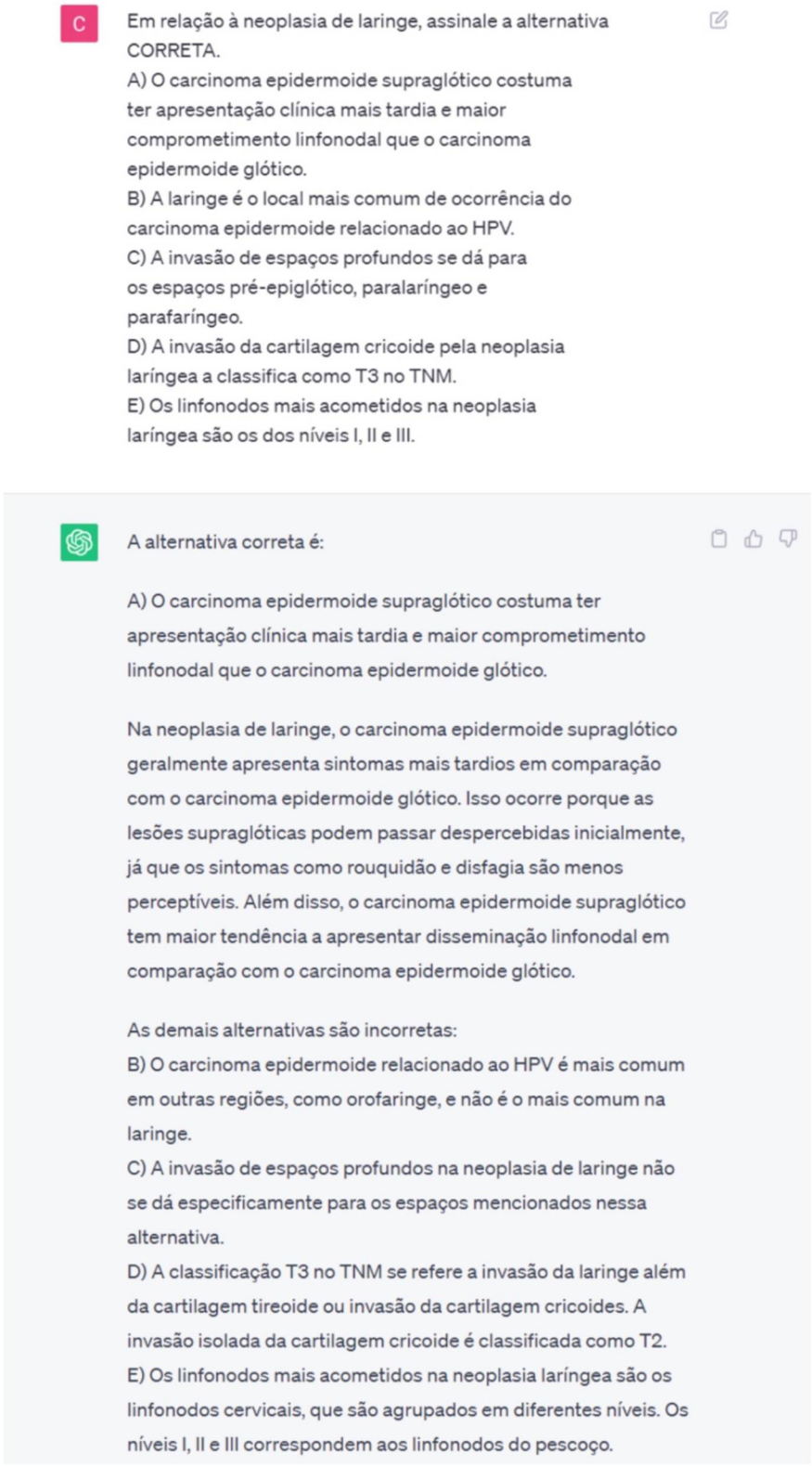
FIGURA 5 – EXEMPLO DE RESPOSTA CORRETA (ORDEM INFERIOR)




FONTE: ELABORADA PELOS AUTORES (2024).

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

FIGURA 6 – EXEMPLO DE RESPOSTA CORRETA (ORDEM SUPERIOR)

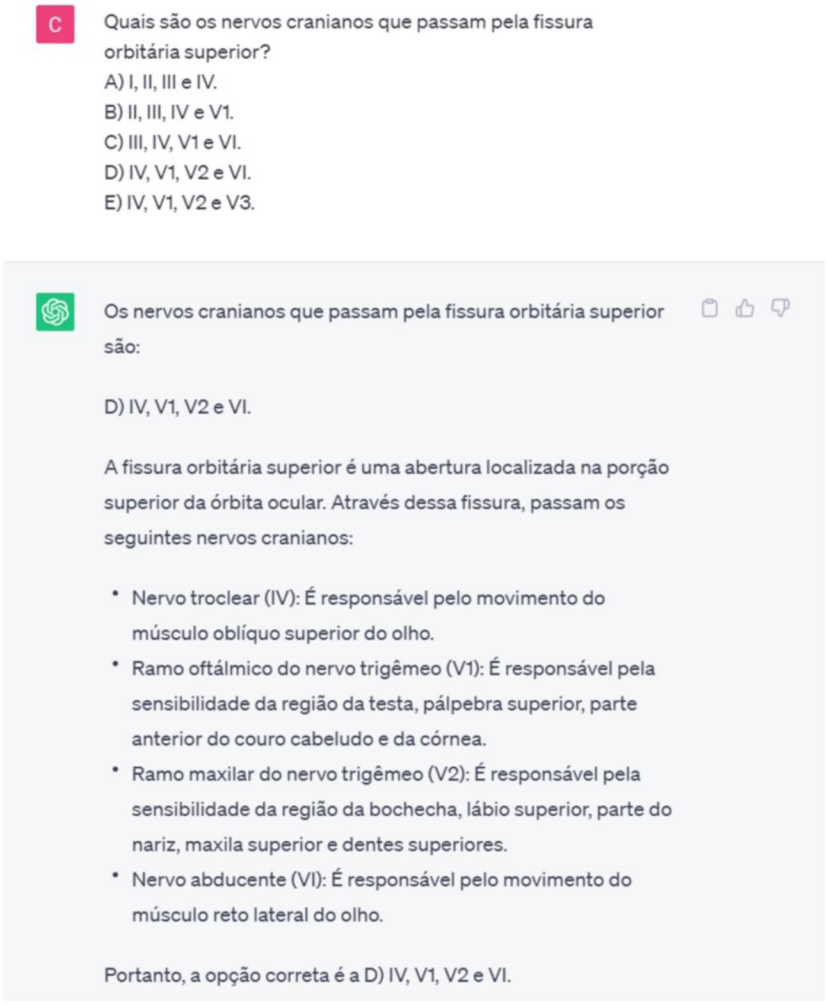



FONTE: ELABORADA PELOS AUTORES (2024).

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

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 7 mostra um exemplo de erro de uma questão a respeito de anatomia, a FIGURA 8 mostra um exemplo de acerto em questão que envolve descrição de achados e a FIGURA 9 mostra um exemplo de acerto em questão que envolve manejo clínico.

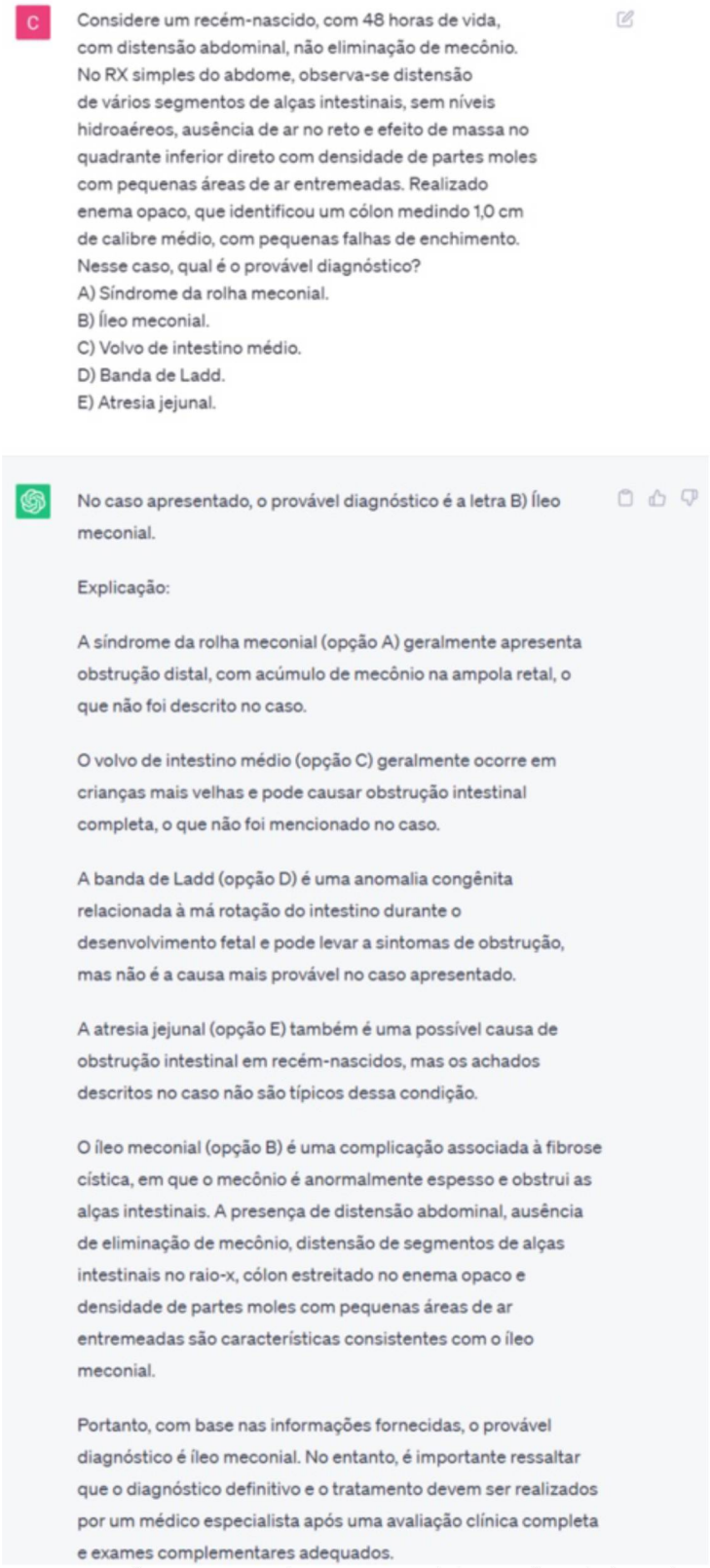
FIGURA 7 – EXEMPLO DE RESPOSTA INCORRETA




FONTE: ELABORADA PELOS AUTORES (2024).

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

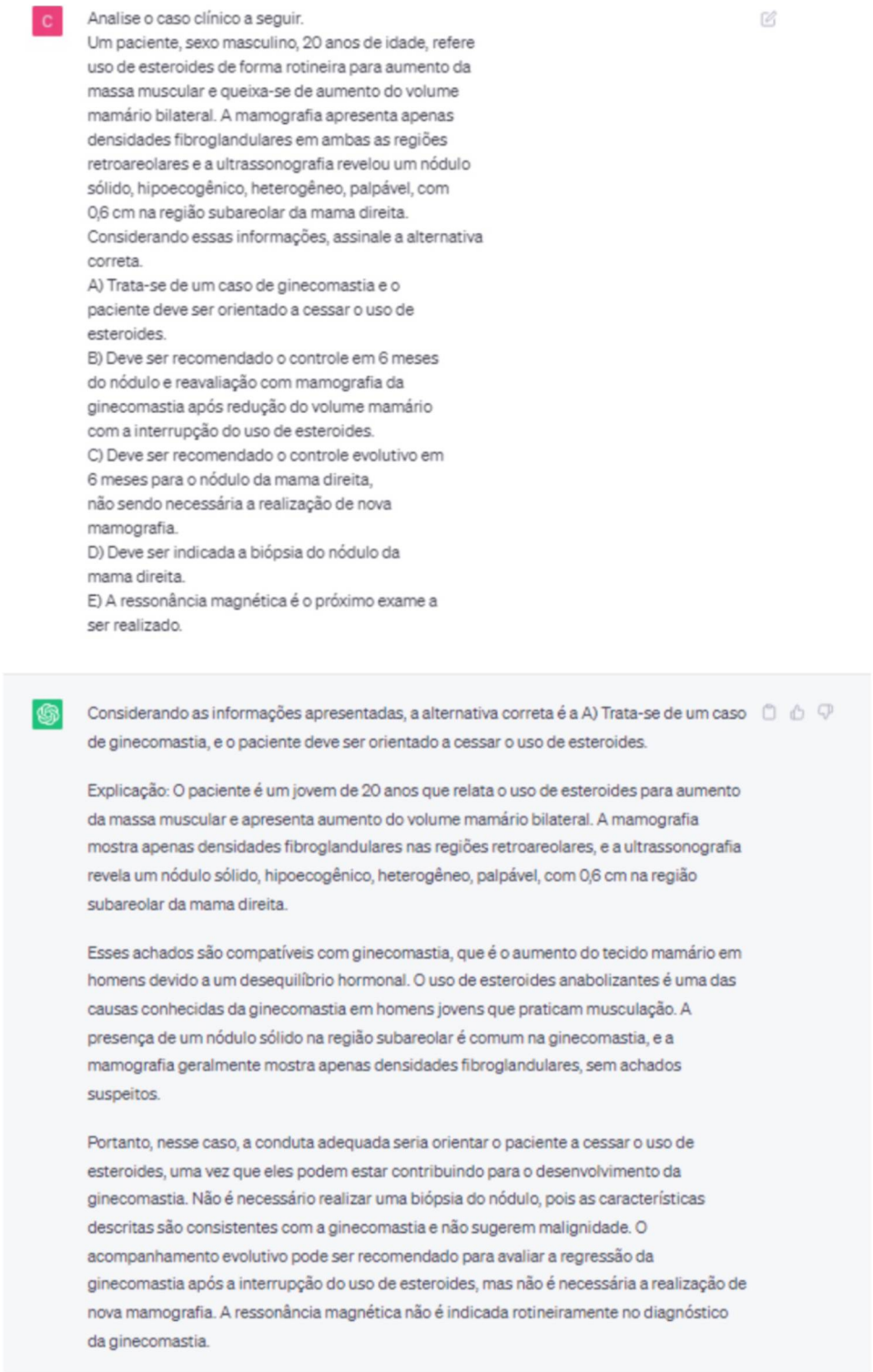
FIGURA 8 – EXEMPLO DE RESPOSTA CORRETA EM DESCRIÇÃO DE ACHADOS




FONTE: ELABORADA PELOS AUTORES (2024).

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

FIGURA 9 – EXEMPLO DE RESPOSTA CORRETA EM MANEJO CLÍNICO




FONTE: ELABORADA PELOS AUTORES (2024).

LEGENDA: 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.

### 4.2.2. Questões de residência médica

A acurácia em questões que avaliam habilidades cognitivas de ordem inferior (63,8%, 99 de 155) foi superior àquela vista nas questões que avaliam habilidades cognitivas de ordem superior (52,5%, 181 de 345) (P= 0,01).

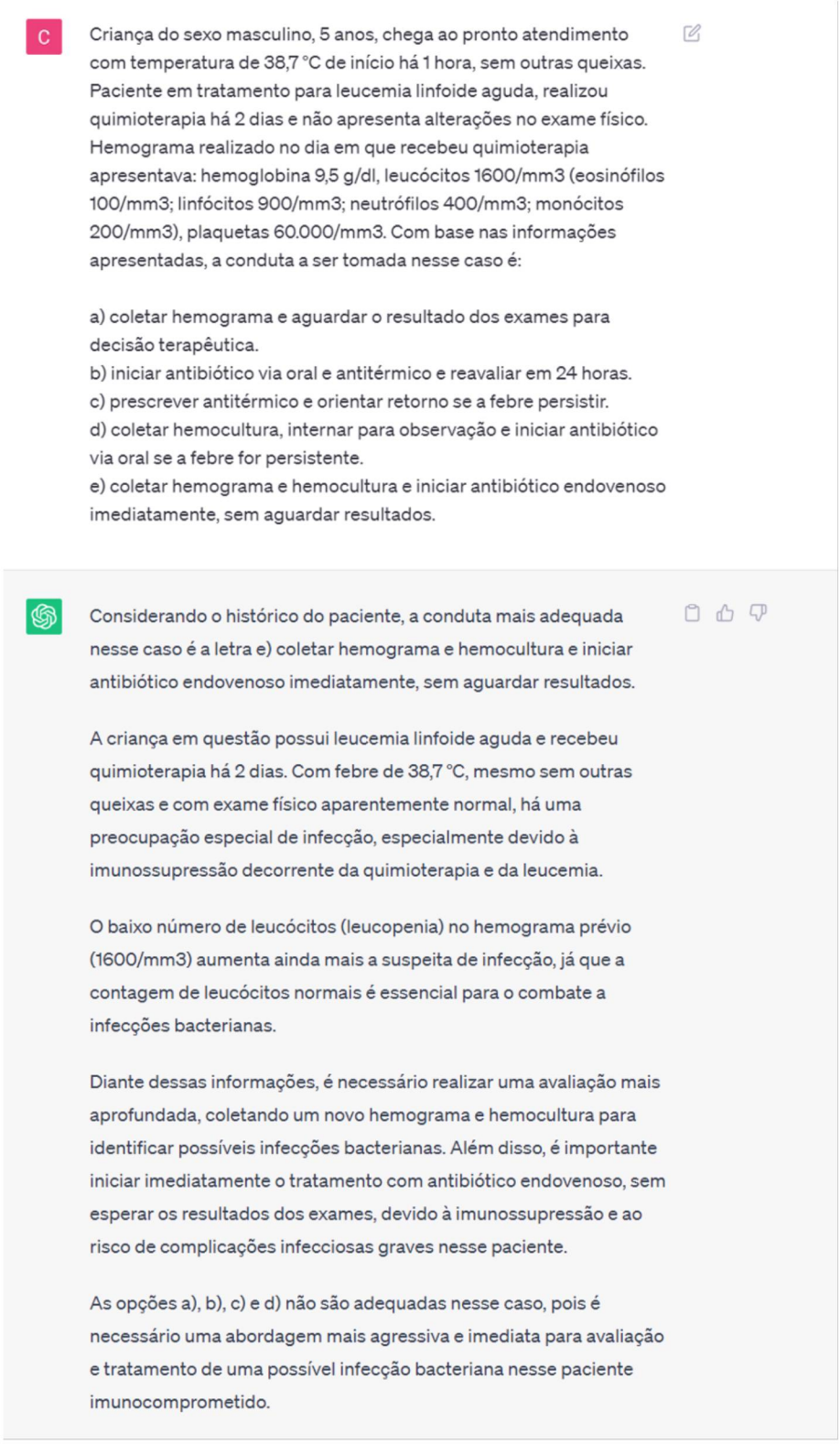
A performance de acordo com o estilo de questão está sintetizada na TABELA 4, não tendo havido diferença estatisticamente significativa entre os variados estilos (P < 0.01). As FIGURAS 10 a 12 mostram exemplos de respostas corretas do ChatGPT a esses diferentes tipos de questões.

TABELA 4 – DESEMPENHO DO CHATGPT DE ACORDO COM O ESTILO DE QUESTÃO

| **Tipo de questão** | **N total** | **Acertos** | **Acurácia (%)** |
|---|---|---|---|
| Descrição | 19 | 14 | 74 |
| Manejo clínico | 121 | 65 | 54 |
| Aplicação de conceito | 133 | 69 | 52 |
| Cálculo ou classificação | 26 | 16 | 61 |
| Revisão teórica do tema | 116 | 57 | 49 |
| Indicar o diagnóstico | 56 | 38 | 68 |
| Indicar o tratamento | 29 | 21 | 72 |

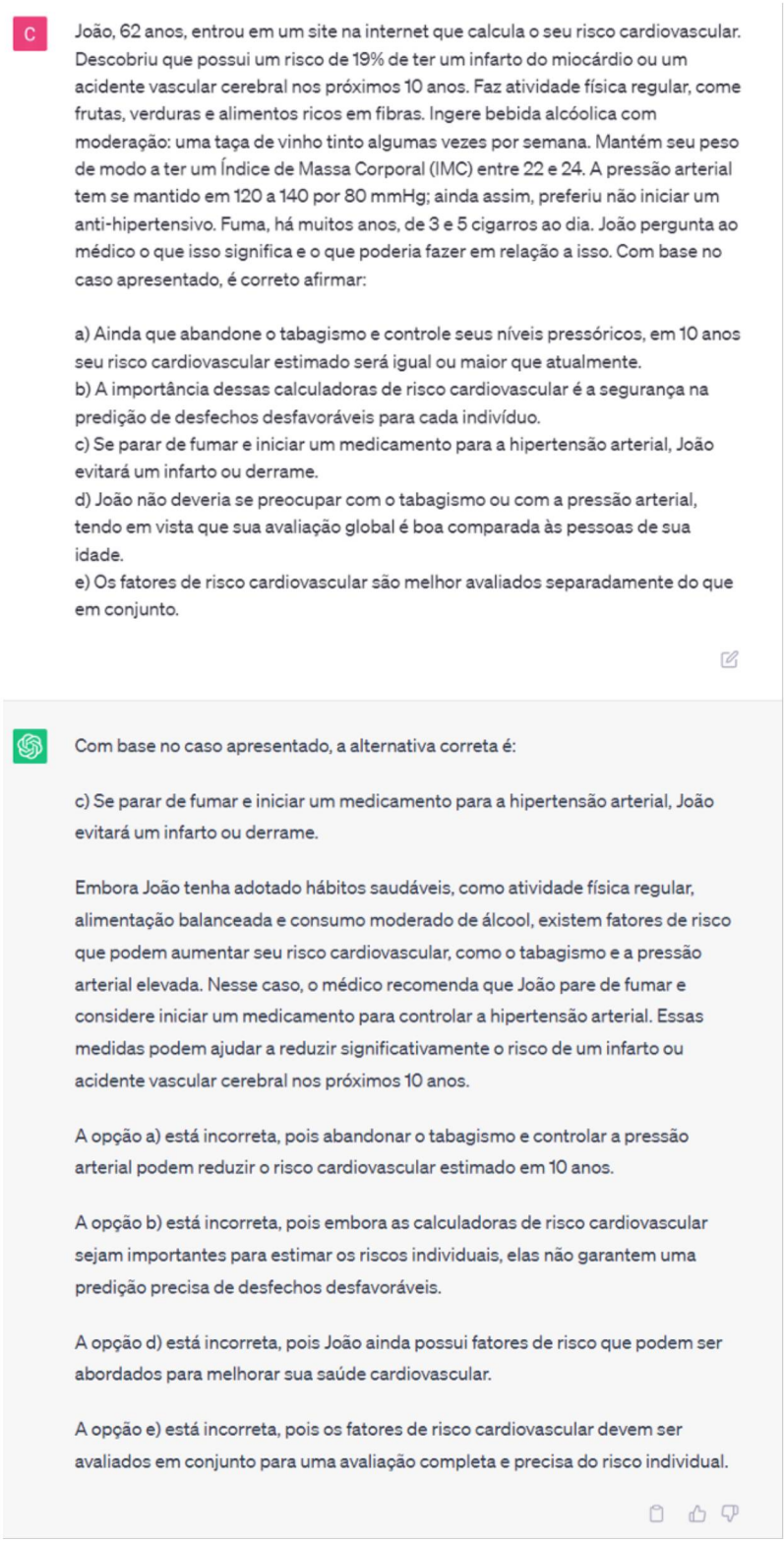
FONTE: ELABORADA PELOS AUTORES (2023).

FIGURA 10 – EXEMPLO DE UMA PERGUNTA DE ORDEM SUPERIOR NO ESTILO “MANEJO CLÍNICO” (INDICAR A CONDUTA EM UM QUADRO CLÍNICO).




FONTE: ELABORADA PELOS AUTORES (2024).

FIGURA 11 – EXEMPLO DE UMA PERGUNTA DE ORDEM SUPERIOR NO ESTILO “CÁLCULO OU CLASSIFICAÇÃO”.




FONTE: ELABORADA PELOS AUTORES (2024).

LEGENDA: Neste caso, a resposta correta seria a letra “a”.

## 4.3. DESEMPENHO POR GRANDE ÁREA

### 4.3.1. Desempenho por grande área da radiologia

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).

### 4.3.2. Desempenho por grande área da medicina

A maior acurácia do ChatGPT foi obtida nas questões de clínica médica (70%), seguida por pediatria (57%), cirurgia geral (56%), ginecologia e obstetrícia (51%) e medicina preventiva e social (46%). A análise estatística não revelou diferença significativa entre essas áreas (p>0.05).

## 4.4. DESEMPENHO POR ANO DA QUESTÃO EM RADIOLOGIA

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).

## 4.5. 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 5 a 12) também chama a atenção, mesmo em questões em que sua resposta foi a incorreta (FIGURA 3).

Outro fato interessante é de que na maior parte das questões (65% das de radiologia, 107 de 165, e 67% das de residência, 338 de 500), não apenas o programa indicou a resposta correta, mas também analisou todas as assertivas, indicando porque as julgou como incorretas (Figuras 1, 2, 4 e 7).

## 5 DISCUSSÃO

Neste estudo, 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 questões nacionais foi pior do que o observado em questões norte-americanas da mesma especialidade - 53,3% naquelas e 69% nestas - (BHAYANA; KRISHNA; BLEAKNEY, 2023), o que pode estar relacionado a diferenças entre as provas de acordo com o conhecimento específico que cada país exige dos futuros radiologistas ou a um possível papel da diferença entre as línguas (pois é provável que o ChatGPT tenha mais dados em língua inglesa em sua base de dados, o que lhe conferiria mais familiaridade com essa língua e facilitaria a interpretação das questões). Novos estudos similares realizados em outros países poderão ajudar a compreender tais diferenças. A tabela 5 compara o desempenho do ChatGPT nas perguntas utilizadas neste estudo e em outros estudos similares realizados em outros países.

TABELA 4 – DESEMPENHO DO CHATGPT 3.5 EM DIFERENTES TESTES

| **Estudo** | **Questões apresentadas** | **Acertos (%)** |
|---|---|---|
| Leitão et al., 2023 | Avaliação anual de residentes do CBR | 53 |
| Leitão et al., 2023 | Provas de residência - UFPR | 56 |
| Bhayana et al., 2023 | Prova de título em radiologia (EUA) | 69 |
| Almeida et al., 2023 | Prova de título do CBR | 63 |
| Weng et al., 2023 | Prova de título em medicina de família (Taiwan) | 42 |
| Ali et al., 2023 | Prova de título em neurocirurgia (EUA) | 66 |
| Kung et al., 2023 | Avaliação anual de residentes em ortopedia dos EUA | 54 |
| Gilson et al., 2023 | Prova para obtenção da licença médica nos EUA (step 1) | 64 |

| | | |
|---|---|---|
| Gilson et al., 2023 | Prova para obtenção da licença médica nos EUA (step 2) | 58 |
| Jung et al., 2023 | Prova para obtenção da licença médica na Alemanha (etapa 1 – pré-clínica) | 60 |
| Jung et al., 2023 | Prova para obtenção da licença médica na Alemanha (etapa 2 – especialidades médicas) | 67 |

FONTE: ELABORADA PELOS AUTORES (2024).

A análise das 77 questões de radiologia 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. Estas últimas consistem em uma das principais preocupações envolvendo o uso de modelos de linguagem grande, pois, quando confrontada com informações com as quais não está completamente familiarizada, é possível que a máquina anda assim gere respostas críveis, porém incorretas (SHEN at al., 2023). A ausência de alucinações neste estudo, contudo, está de acordo com a tendência vista na literatura, uma vez que alucinações não são tão frequentes em *chatbots* porque eles 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, não para gerar novas informações (ALKAISSI; MCFARLANE,2023), o que costuma ser a fonte de alucinações. Estudo semelhante recente também confirmou essa tendência (PATIL et al., 2023) 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 (BHAYANA; KRISHNA; BLEAKNEY, 2023), tendo sido reforçado neste estudo. Este achado mostra a capacidade da 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 neste 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. É possível, inclusive, que o maior número de questões de física entre as perguntas apresentadas a todos os residentes (desde o primeiro ano) tenha influenciado o maior número de acertos do chatGPT nessas perguntas do que nas perguntas para residentes de segundo e terceiro ano. Os modelos de linguagem grande, 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.

Apesar de o ChatGPT ter apresentando melhor desempenho em algumas áreas da medicina e da radiologia (como clínica médica e radiologia abdominal) do que em outras (como medicina preventiva e ultrassonografia), não houve diferença estatisticamente significativa neste estudo, o que sinaliza que essa diferença pode ser decorrente apenas do acaso. É possível que novos estudos com maior quantidade de questões, e após maior treinamento da máquina com dados referentes à radiologia, venha a mostrar alguma inclinação a uma área do conhecimento que ainda não está clara. Atualmente, a ausência de diferença estatisticamente significativa entre as subespecialidades da radiologia e as diferentes áreas da medicina também pode ser compreendida pela baixa familiaridade do ChatGPT com os termos e jargões que compõem cada uma delas. A medicina e cada uma das suas subáreas, inclusive a radiologia, possuem um 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 levada 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 daquele 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 medicina ou radiologia como um todo, mas também em suas subáreas.

Outro achado digno de nota deste estudo é o fato de o ChatGPT ter analisado todas as assertivas da maior parte das questões que lhe foram apresentadas. Não ficou claro neste estudo 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 acadêmicos de medicina e residentes que desejem utilizar as questões de provas antigas disponibilizadas pelo CBR e pelas comissões de residência médica 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%) (BHAYANA; KRISHNA; BLEAKNEY, 2023). 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, naquele, por pesquisadores de um único centro), poderiam contribuir para essa divergência. Além disso, embora ainda não esteja claro, novamente é possível que o idioma também exerça alguma influência na performance do ChatGPT. A ferramenta não traduz diretamente as perguntas da língua apresentada para o inglês, mas utiliza a familiaridade com ela (obtida durante a formação de sua base de dados, que foi enriquecida com textos em diferentes idiomas) para respondê-la no idioma original. Uma vez que há maior disponibilidade de literatura em língua inglesa para treinamento do modelo, teoricamente, haverá menor familiaridade com perguntas em português, o que pode afetar o desempenho final. Além disso, a máquina pode não

compreender 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.

Um estudo similar foi publicado por outra equipe brasileira no final de 2023 (ALMEIDA et. al, 2023), obtendo-se um desempenho um pouco melhor, com a diferença de que naquele estudo foram utilizadas questões da prova de título de radiologia do CBR, enquanto, neste, questões da avaliação anual de residentes. Outra diferença é naquele estudo os autores forneceram diferentes *prompts* ao ChatGPT antes de apresentaram as perguntas a ele (como, por exemplo, "escolha apenas a alternativa que representa a resposta correta"), algo que não foi feito neste estudo, em que se objetivou verificar o desempenho espontâneo do chatGPT, sem *prompts* específicos. É possível que essas diferenças tenham influenciado o resultado diferente obtido entre os autores, porém novos estudos utilizando *prompts* ou novas questões elaboradas pelo CBR podem ser necessários para elucidar essa diferença de desempenho. Ainda não há estudos publicados utilizando questões de residência médica, como neste estudo.

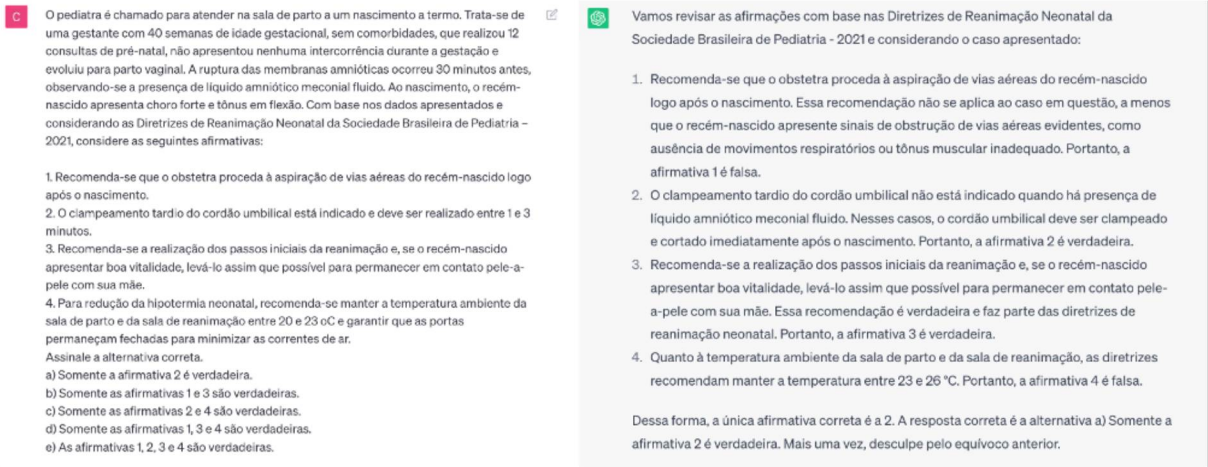
Há ainda pelo menos dois outros aspectos que podem explicar os erros apresentados pelo ChatGPT. O primeiro seria a presença de frases e termos (chamados em linguagem de computação de "lemas") nas questões que são similares, ou até mesmo redundantes, a lemas que ativam outra via de associação no algoritmo. Um exemplo está ilustrado na FIGURA 11, em que a frase "evitará um novo infarto" fez sentido para o algoritmo, mesmo que não seja a resposta correta, pois as mudanças de estilo de vida presentes na assertiva ativam as vias de associação que levam à conclusão de que haverá prevenção de eventos cardiovasculares.

Outro fator relacionado a discrepância entre a performance real e a previamente estabelecida de um software como o ChatGPT é relacionado a um mecanismo de análise matemática chamado de *Overfitting* (TILLMANN, 2014). *Overfitting* ocorre quando o modelo é treinado e apresenta uma performance boa com sua base de treinamento (com a qual ele foi gerado), porém com dificuldades de generalização na aplicação real, como nas questões de residência. Isso ocorre em modelos complexos como o próprio chatGPT, ou por um período de treinamento prolongado, que levam o algoritmo a "memorizar" os dados de treinamento, mas sem conseguir aplicá-los corretamente no cenário real.

A respeito das questões de provas de residência, o desempenho obtido nas questões nacionais foi similar a estudos semelhantes realizados no exterior (ALI et al.; GILSON et al., 2023; WENG et al., 2023). Esses resultados podem ser considerados animadores, pois mostram que ainda há espaço para que a IA seja treinada em conhecimento médico e, uma vez enriquecidos nessa área, os programas provavelmente alcançarão desempenhos ainda melhores. Ainda se discute quais impactos essa nova realidade trará sobre o exercício da medicina (KING et al., 2023; VALENTE et al., 2023) apontando as principais evidências para seu potencial uso diagnóstico, terapêutico e de vigilância em saúde (KING et al., 2023). Seu papel na educação médica, contudo, parece especialmente promissor.

O ChatGPT também possui razoável poder de síntese, sendo capaz de sumarizar informações disponíveis em consensos e diretrizes estabelecidas por sociedades de especialistas. Após ter sido alimentada com os artigos e consensos a respeito de um tema específico, é possível pedir à inteligência artificial uma síntese das informações neles contidas de forma rápida e eficiente, dispensando a busca e revisão de todos os arquivos individualmente, o que demandaria mais tempo do estudante. Um exemplo desta habilidade foi visto neste estudo, quando, após ter sido corrigido ao errar uma questão, o ChatGPT foi capaz de acertá-la citando as diretrizes nacionais que estavam sendo referenciadas (FIGURA 12). Essa habilidade é valiosa, pois comprova que a plataforma responde às perguntas com base em evidências, como todo estudo na medicina deve se basear. Isso mostra também que tal ferramenta já caminha no sentido de auxiliar os estudantes a compreenderem que a tomada de decisões médicas deve ser sempre baseada em evidências e de aproximá-los das principais referências bibliográficas que devem ser conhecidas.

FIGURA 12 - EXEMPLO DE REFERÊNCIA A UMA DIRETRIZ




FONTE: ELABORADA PELOS AUTORES (2024).

Este estudo apresenta limitações. Foram utilizadas apenas questões objetivas teóricas que não envolviam a interpretação de imagens, pois o ChatGPT, em sua versão 3.5, 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.

## 6 CONCLUSÃO

Em resumo, o ChatGPT apresentou um desempenho similar ao de candidatos reais frente a questões nacionais de concursos de residência médica e em questões para avaliar residentes brasileiros de radiologia. Não houve diferença estatisticamente significante em seu desempenho comparando as grandes áreas da medicina ou as subespecialidades da radiologia entre si. A análise de suas respostas sugere um possível papel auxiliar dessa ferramenta para acadêmicos de medicina, médicos que se preparam para concursos de residência e residentes que se preparam para avaliações estudarem os assuntos abordados, mas ainda é necessário que o *chatbot* seja treinado com mais dados específicos da medicina. A comunidade acadêmica

deve permanecer atenta ao ChatGPT e a outros modelos de linguagem grande para aproveitar essas potencialidades.

# REFERÊNCIAS

ADAMS, L.C. et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. **Radiology**, v. 307, n. 4, 2023.

AHMED, A. et al. The Effectiveness of Wearable Devices Using Artificial Intelligence for Blood Glucose Level Forecasting or Prediction: Systematic Review. **J Med Internet Res**, v. 14,n. 25, 2023.

ALI, R.; TANG, O.Y.; CONNOLLY, I.D.; SULLIVAN, P.L.Z; SHIN, Z.H.; FRIDLEY, J.S. et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations.**medRxiv**, v.3,n.25, 2023.

ALKAISSI, H.; MCFARLANE, S.I. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. **Cureus**, v. 15, n.2, 2023.

ALMEIDA, L.C.; FARINA, E.M.J.M.; KURIKI, P.E.A.; ABDALA, N.; KITAMURA, F.C. Performance of ChatGPT on the Brazilian Radiology and Diagnostic Imaging and Mammography Board Examinations. **Radiology: Artificial Intelligence**,v.0, n.0:ja

ARDILLA, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. **Nat Med**, v. 25, p. 954–961, 2019.

BHAYANA, R.; KRISHNA, S,; BLEAKNEY, RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. **Radiology**, 2023.

BISWAS, S. ChatGPT and the Future of Medical Writing. **Radiology**, v. 307, n.2, 2023.

BLOOM, B.S. et al. **Taxonomy of educational objectives: The classification of educational goals.** Vol. Handbook I: Cognitive domain. New York: David McKay Company, 1956. Disponível em: https://eclass.uoa.gr/modules/document/file.php/PPP242/Benjamin%20S.%20Bloom%20-%20Taxonomy%20of%20Educational%20Objectives%2C%20Handbook%201_%20Cognitive%20Domain-Addison%20Wesley%20Publishing%20Company%20%281956%29.pdf. Acesso em: 23 jan.2024.

CHADEBECQ, F., LOVAT, L.B.; STOYANOV, D. Artificial intelligence and automation in endoscopy and surgery. **Nat Rev Gastroenterol Hepatol**, v. 20, p. 171–182, 2023.

CHAN, K.; ZARY, N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. **JMIR Med Educ**, v. 5, n.1, 2019.

COLÉGIO BRASILEIRO DE RADIOLOGIA E DIAGNÓSTICO POR IMAGEM. Avaliação Anual de Residentes – Provas anteriores. Disponível em: https://cbr.org.br/avaliacao-anual-de-residentes-provas-anteriores/. Acesso em: 02 jun

2023.

DEVLIN, J.; CHANG, M.W.; LEE, K.; TOUTANOVA, K. BERT. **Pre-training of Deep Bidirectional Transformers for Language Understanding**. Disponível em: https://arxiv.org/abs/1810.04805/. Acesso em: 20 nov. 2023.

DREYFUS, S.E. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. **Journal of Guidance, Control, and Dynamics**, v. 13, n. 5, 1990.

EPSTEIN R. Assessment in medical education. **N Engl J Med**, v. 25, n. 356, n.4, p. 387-396, 2007.

FRIZE, M.; FRASSON, C. Decision-support and intelligent tutoring systems in medical education. **Clin Invest Med**, n. 23, v. 4, p. 266-269, 2000.

GARNELO, M.G.; SHANAHAN, M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. **Current Opinion in Behavioral Sciences**, v. 29, p. 17-23, 2019.

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 Med Educ**, v.12, n.3, 2023.

HENDLER, J. Avoiding another AI Winter. **IEEE Intelligent Systems**. v. 23, n.2, p. 2–4, 2008.

HEWSON, M.; LITTLE, M. Giving feedback in medical education: verification of recommended techniques. **J Gen Intern Med**, v. 13, n. 2, p.111-116, 1998.

IRMAN, AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. **Inform Med Unlocked**, v. 20, 2020.

JUNG, L.B., et al. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. **Dtsch Arztebl** Int, v. 120, n. 21,p. 373-374, 2023.

KAUL, V.; ENSLIN, S.; GROSS, S.A. History of artificial intelligence in medicine. **Gastrointestinal Endoscopy**, v. 92, n. 4, p. 807-812, 2020.

KIM, M. et al. Deep Learning in Medical Imaging. Neurospine, v. 16, n. 4, p. 657-668, 2019.

KING, M.R. The Future of AI in Medicine: A Perspective from a Chatbot. **Ann Biomed Eng**, v.51, p. 291-295, 2023.

KITAMURA, F.C. ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment. **Radiology**, v. 307, n.2, 2023.

LINDEGGER, D.J.; WAWRZYNSKI, J.; SALEH, G. M. Evolution and Applications of Artificial Intelligence to Cataract Surgery. **Ophthalmol Sci**, v. 25, n. 2, 2022.

LOCKE, S. et al. Natural language processing in medicine: A review. **Trends in Anaesthesia and Critical Care**,v. 38, p. 4-9, 2021.
KOOHY, H. The rise and fall of machine learning methods in biomedical research. **F1000Res**, v.2, n. 6, 2012.

KUNG, J.E. et al. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. **JBJS Open Access**, v.8, n. 3, 2023.

MANYIKA, J. **An overview of Bard: an early experiment with generative AI**. Disponível em: https://ai.google/static/documents/google-about-bard.pdf. Acesso em: 24 jan. 2024.

MCCARTHY, J.et al. **A proposal for the dartmouth summer research project on artificial intelligence**. Disponível em: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf. Acesso em: 23 jan. 2024.

MCFADDEN P.; CRIM, A. Comparison of the Effectiveness of Interactive Didactic Lecture Versus Online Simulation-Based CME Programs Directed at Improving the Diagnostic Capabilities of Primary Care Practitioners. **J Contin Educ Health Prof**, v. 36, n.1, p.32-37.

MOOR, J. The Dartmouth College Artificial Intelligence Conference: The Next Fifty years. **AI Magazine**, v. 27, n. 4, p.87–89, 2006.

MORREEL, S; MATHYSEN D, VERHOEVEN V. ChatGPT passes multiple-choice family medicine exam. **Med Teach**, v.45, n.6, p.665-666.

NADKARNI, P.M.; OHNO-MACHADO, L.; CHAPMAN, W.W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, v. 18, n. 5, p. 544-551, 2011.

NÚCLEO DE CONCURSOS. Disponível em: https://rb.gy/1nr01. Acesso em 31 mai 2023.

OPENAI. **OpenAI Gym Beta (2016).** Disponível em: https://openai.com/research/openai-gym-beta. Acesso em: 22 jan. 2024.

OPENAI. **OpenAI Five (2018).** Disponível em: https://openai.com/research/openai-five. Acesso em: 22 jan. 2024.

OPENAI. **GPT-2: 1.5B release (2019).** Disponível em: https://openai.com/research/openai-gym-beta. Acesso em: 22 jan. 2024.

OPENAI. **Introducing ChatGPT**. Disponível em: http://www.impactscan.org/. Acesso em: 02 jun. 2023.

OPENAI. **GPT-4**. Disponível em: https://openai.com/research/gpt-4. Acesso em: 22 jan. 2024.

PALIHAPITIYA, C. **Quick Essay: A Short History of OpenAI**. Disponível em: https://chamath.substack.com/p/a-short-history-of-openai. Acesso em: 22 jan. 2024.

PAN, X. et al. The artificial intelligence-based model ANORAK improves histopathological grading of lung adenocarcinoma. **Nat** Cancer, 2024.

PATIL, N.S.; HUANG, R.S.; VAN DER POL C.B; LAROCQUE, N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. **Canadian Association of Radiologists Journal**, 2023.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. **Improving language understanding by generative pre-training**. Disponível em: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Acesso em: 02 jun. 2023.

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine Learning in Medicine. **N Engl J Med**, v. 380, n. 14, p.1347-1358, 2019.

SARRAYRIH, M; ILYAS, M. Challenges of online exam, performances and problems for online university exam. **International Journal of Computer Science Issues**, v. 10, n. 1, p. 439, 2013.

SOCIEDADE BRASILEIRA DE PEDIATRIA. **Programa de Reanimação Neonatal. Recomendações para assistência ao recém-nascido na sala de parto**. Disponível em: www.sbp.com.br/reanimacao. Acesso em 04 jun 2023.

SHEN, Y. et al. ChatGPT and Other Large Language Models Are Double-edged Swords. **Radiology**, v. 307, n.2, 2023.

TAMKIN, A.; BRUNDAGE, M.; CLARK, J.; GANGULI, D. **Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models**. Disponível em: https://arxiv.org/abs/2102.02503. Acesso em: 20 nov. 2023.

THIRUNAVUKARASU, A.J.; TING, D.S.J.; ELANGOVAN, K. et al. Large language models in medicine. **Nat Med**, v. 29, p. 1930–1940, 2023.

TILLMANN, A.M. On the Computational Intractability of Exact and Approximate Dictionary Learning». **IEEE Signal Processing Letters**, v.22, n.1, p.45-49, 2014.

TURING, A.M. Computing machinery and intelligence. **Mind**, v.59, n.236, p.433-460, 1950.

VALENTE, M.; BELLINI, V.; DEL RIO, P.; FREYRIE, A; BIGNAMI, E. Artificial Intelligence Is the Future of Surgical Departments.... Are We Ready? **Angiology**, v. 74, n.4, p.397-398, 2023.

VETTORETTI M., et al. Advanced Diabetes Management Using Artificial Intelligence and Continuous Glucose Monitoring Sensors. **Sensors (Basel)**, v. 20, n. 14, 2020.

WANG, F.; PREININGER, A. AI in health: state of the art, challenges, and future directions. **Yearbook of medical informatics**, v.28, p.16–26, 2019.

WARTMAN, S.; COMBS, C. Medical Education Must Move From the Information Age to the Age of Artificial Intelligence. **Acad Med**, v. 93, n. 8, p. 1107-1109, 2018.

WENG, T.L.; WANG, Y.M.; CHANG, S.; CHEN, T.J.; HWANG, S.J. ChatGPT failed Taiwan's Family Medicine Board Exam. **J Chin Med Assoc**, 2023.

WRITESONIC. **Chatsonic – Best ChatGPT Alternative for Content Creation.** Disponível em: https://writesonic.com/chat. Acesso em 23 jan. 2024.

# APÊNDICE 1 - CARTA DE ACEITE DO PRIMEIRO ARTIGO PARA PUBLICAÇÃO EM REVISTA QUALIS A4

Decision on Manuscript - Artigo RB-2023-0083.R1 Caixa de entrada ×

**Radiologia Brasileira** <onbehalfof@manuscriptcentral.com> 15 de set. de 2023, 21:07
para mim

15-Sep-2023

Prezados Autores,

Referente ao artigo: Desempenho do ChatGPT nas questões da avaliação anual de residentes do Colégio Brasileiro de Radiologia

Código de fluxo: RB-2023-0083.R1

Temos o prazer de informar que o manuscrito acima citado foi aprovado pelo Conselho Editorial e será incluído na programação para publicação na Revista Radiologia Brasileira. Lembramos que algumas modificações poderão ser solicitadas até a publicação do artigo.

Além disso, é necessário o envio de dois documentos, que seguem anexos, antes da publicação do artigo: a Divulgação de Potencial Conflito de Interesses pelo Autor (assinado apenas pelo autor correspondente) e o Termo Copyright (assinado por todos os autores).

Obrigado por submeter seu trabalho e aguardamos novas contribuições.

Atenciosamente,

Prof. Valdair Muglia
Editor-chefe, Radiologia Brasileira

# APÊNDICE 2 - SUBMISSÃO DO SEGUNDO ARTIGO PARA PUBLICAÇÃO EM REVISTA QUALIS B1 (AGUARDA DECISÃO FINAL APÓS REVISÃO)

Revista Brasileira de Educação Médica - Manuscript ID RBEM-2023-0162 Caixa de entrada ×

**Revista Brasileira de Educação Médica RBEM** <onbehalfof@manuscriptcentral.com> 29 de jun. de 2023, 17:32
para mim, glucca11, ledamrabelo2, dante.escuissato

Traduza para o português

29-Jun-2023

Dear Dr. LEITAO:

Your manuscript entitled "Desempenho do ChatGPT em questões de residência: reflexões sobre a inteligência artificial na educação médica" has been successfully submitted online and is presently being given full consideration for publication in the Revista Brasileira de Educação Médica.

Your manuscript ID is RBEM-2023-0162.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your street address or e-mail address, please log in to ScholarOne Manuscripts at https://mc04.manuscriptcentral.com/rbem-scielo and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Center after logging in to https://mc04.manuscriptcentral.com/rbem-scielo.

Thank you for submitting your manuscript to the Revista Brasileira de Educação Médica.

Sincerely,
Revista Brasileira de Educação Médica Editorial Office

## Manuscripts with Decisions

| ACTION | STATUS | ID | TITLE | SUBMITTED | DECISIONED |
|---|---|---|---|---|---|
| a revision has been submitted (RBEM-2023-0162.R1) | Contact Journal<br>ADM: RBEM, Revista Brasileira de Educação Médica<br>• Minor Revision (01-Dec-2023)<br>• a revision has been submitted<br>view decision letter | RBEM-2023-0162 | Desempenho do ChatGPT em questões de residência: reflexões sobre a inteligência artificial na educação médica<br>View Submission | 09-Jul-2023 | 01-Dec-2023 |

# APÊNDICE 3 - PRIMEIRO ARTIGO

**RB-2023-0083 – Original Article**

## 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, Brazil.

Correspondence: Dr. Cleverson Alex Leitão. Universidade Federal do Paraná. Rua General Carneiro, 181, Alto da Glória. Curitiba, PR, Brazil, 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.

Submitted 13 July 2023. Revised 29 August 2023. Accepted 15 September 2023.

**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). Fisher's exact test, the chi-square test, and analysis of variance were employed.

**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.

**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, 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 ano (R1, R2 ou R3). Foram realizados testes exato de Fisher, qui-quadrado e ANOVA.

**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%; 18/20) do que em questões clínicas (48%; 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.

Running title: Performance of ChatGPT on questions formulated by the CBR

## INTRODUCTION

Artificial intelligence (AI) is the general name given to computing methods that simulate the learning pattern of the human brain[1]. The rapid advances recently made in this field of knowledge have raised questions about how it will impact diverse professions, including that of medicine, in the future. Among the existing AI models, the Chat Generative Pretrained Transformer (ChatGPT) has gained prominence, not only in the scientific literature[2–4] but also in the popular media[5]. It is an AI tool based on the relationships between AI algorithms and

human language, a strategy known as natural language processing, and has been publicly available since November 30, 2022[6]. Its current model is GPT-3.5, a large language model trained on more than 45 terabytes of textual data. Through neural networks, those data give the tool the capacity to analyze texts and generate texts similar to those written by humans[7]. Although it has not been specifically trained for medical use, studies have demonstrated its promising role in medical practice[8] and in academic medical writing[9]. As a way of evaluating the knowledge of ChatGPT on medical topics, its performance has been tested on academic examinations that evaluate real students, such as the test for obtaining a medical license in the United States[10], and on questions for obtaining specialist degrees in radiology in Canada and the United States[7], as well as on those for obtaining a degree in family medicine in Taiwan[11], with results that show its performance to be, in general, close to that required for approval.

In the specific context of radiology, AI has been used mainly as an aid in image interpretation, although language models such as ChatGPT have also shown potential as an aid in writing radiological reports[12] and in clinical decision making[4]. A better understanding of the performance of AI in the context of problems encountered in daily radiology practice can help us understand how it will influence the future of the profession. With that objective in mind, we sought to evaluate the performance of ChatGPT on questions prepared by the *Colégio Brasileiro de Radiologia* (CBR, Brazilian College of Radiology) for the annual evaluation of residents in radiology and diagnostic imaging, analyzing its answers to determine what its current strengths and weaknesses are.

## MATERIALS AND METHODS

This was a prospective analytical study carried out between May 24 and June 3 of 2023. Because the study did not involve human beings or patient data, approval by an institutional review board was not required.

### Questions for the annual evaluation of radiology residents

A total of 165 questions were selected from the annual evaluation tests for residents in radiology and diagnostic imaging applied by the CBR in the years 2018, 2019, and 2022, which are available online for public access on the CBR website[13] and whose use has been authorized by the CBR Committee for Certification and Licensing. All questions were of the multiple-choice type, with only one correct answer and four incorrect answers. Questions with images were excluded, because ChatGPT does not yet have the ability to interpret images. They were divided according to their topic into physics questions (n = 20) and clinical questions (n = 145), the latter representing the main fields of knowledge and subspecialties of radiology: abdominal

imaging (n = 20); thoracic imaging (n = 15); breast imaging (n = 15); neuroradiology (n = 15); pediatric radiology (n = 15); musculoskeletal imaging (n = 15); contrast media (n = 15); ultrasound (n = 15); obstetrics and gynecological imaging (n = 10); and miscellaneous, including positron-emission tomography/computed tomography, densitometry, Doppler ultrasound, and radiation safety (n = 10).

Subsequently, the questions were subdivided, according to the principles of Bloom's taxonomy[14], into questions that assess lower-order cognitive skills (remember an idea, memorize a concept) and questions that assess higher-order cognitive skills (evaluate, analyze, synthesize knowledge obtained). Those that assess higher-order cognitive skills were again divided, by style, into six subcategories: description of a clinical finding or sign; clinical management of a case; application of a concept; calculation or classification of the findings described; correlations between diseases; and anatomy. Each of the authors, working independently, classified all of the questions. In cases of disagreement, the final classification was obtained by consensus.

Finally, the questions were divided into three tiers: those applied to all residents (n = 92); those applied to second- and third-year residents (n = 34); and those applied to third-year residents only (n = 39).

**ChatGPT**

The most recent version of ChatGPT available (May 24, 2023; OpenAI) was used. Although this tool was trained with more than 45 terabytes of data in text forma (from web pages, books, and scientific articles), those data were not provided specifically to meet the needs of the radiologist. ChatGPT does not perform internet searches; it answers questions using only its own database.

**Data collection and analysis**

The questions and their respective answer choices were presented to ChatGPT sequentially, one by one, exactly as formulated by the CBR, without providing a specific pre-prompt, and its answers were saved in a text file for later analysis by the researchers. For the questions it answered incorrectly, feedback was provided immediately, the error being explained and the correct answer being supplied, in order to analyze the behavior of the model in response to the correction. In addition to the quantitative analysis of the numbers of correct and incorrect answers, the researchers carried out a qualitative group analysis, obtaining a consensus for comments regarding the answers given.

**Statistical analysis**

To analyze the accuracy rate, the ratio between the number of correct answers and the total number of questions was calculated for all categories (overall; high- and low-order questions; and the question subtypes as described above). The final (overall) ratio was converted to a percentage to represent the accuracy rate.

Comparisons between the question groups (low-order vs. high-order cognitive skills; physical vs. clinical; and one style vs. another style) in terms of the accuracy rate were made by using Fisher's exact test or the chi-square test, as appropriate The analysis among subgroups of questions (by topic and target academic year) was performed with analysis of variance. The statistical analysis was performed with Stata software, version 16.0 (Stata Corp LP, College Station, TX, USA), and post-processing was carried out by using the Analyze Data feature of Microsoft Excel 365. Values of $p < 0.05$ were considered statistically significant.

**RESULTS**

**Overall result**

ChatGPT provided a correct answer on 88 of the 165 questions asked, resulting in a score of 53%, which is well below the 70% defined as a passing score by the CBR. Table 1 shows its performance according to the type and topic of the question.

**Performance by question type**

The performance of ChatGPT was better on questions that assess lower-order cognitive skills, for which it provided the correct answer on 38 (64.4%) of the 59 questions, than on questions that assess higher-order cognitive skills, for which it provided the correct answer on only 50 (47.2%) of the 106 questions, and the difference was statistically significant ($p = 0.01$). Figures 1 and 2 show examples of correct answers on questions that assess lower- and higher-order cognitive skills, respectively.

Among the questions that assess higher-order cognitive skills, the performance of ChatGPT was poorer on those related to anatomy, calculation/classification, and correlations between diseases, although there was no statistically significant difference in comparison with the questions on which it performed better, which were those related to the description of findings, clinical management, and application of concepts ($p > 0.05$). Figure 3 shows an example of a ChatGPT error on a question regarding anatomy, Figure 4 shows an example of a correct answer on a question regarding the description of findings, and Figure 5 shows an example of a correct answer on a question regarding clinical management.

**Performance by question topic**

ChatGPT performed better on physics questions than on clinical questions, and the difference was statistically significant ($p = 0.02$). Among the clinical questions, the accuracy rates were highest for the questions on pediatric radiology, abdominal imaging, and thoracic imaging, although there was no statistically significant difference in comparison with the questions on obstetrics/gynecological imaging and ultrasound, for which the accuracy rates were lowest ($p > 0.05$).

**Performance by target academic year**

ChatGPT performed best on the questions applied to all residents, providing a correct answer on 57 (61.9%) of the 92 questions, followed by those applied to second- and third-year residents, for which it provided a correct answer on 17 (50.0%) of the 34 questions and those applied to third-year residents only, for which it provided a correct answer on 14 (36.9%) of the 39 questions. However, there was no statistically significant difference among the categories ($p > 0.05$).

**Qualitative assessment of the answers**

The unanimous assessment of the evaluators was that the performance of ChatGPT was satisfactory, especially given that its database was not developed specifically for use in the field of radiology. The high degree of assertiveness that the model exhibited in providing its answers, never using words that would indicate doubt or hesitation (Figures 1 to 5), even in answers that were incorrect (Figure 3), was also noteworthy. Another interesting finding is that, on 107 (64.8%) of the 165 questions, the model not only indicated the correct answer but also analyzed all of the other answer choices, indicating why it judged them to be incorrect (Figures 1, 2, and 4).

## DISCUSSION

To our knowledge, this is the first study of its type to be carried out exclusively with data related to Brazil. Our findings make it evident that the accuracy of ChatGPT on radiology questions is not yet high enough to obtain the score required for approval on the annual CBR evaluation of residents in radiology and diagnostic imaging. The performance of ChatGPT on questions designed for radiology residents in Brazil was worse than that observed on questions designed for their counterparts in Canada and the USA[7]—53.3% versus 69.0%—which might be attributable to differences between the two tests in terms of the specific knowledge that each

country demands from its future radiologists. New, similar studies carried out in other countries might clarify such differences.

The analysis of the 77 questions that ChatGPT got wrong shows that its errors can basically be attributed to a lack of knowledge of the subject being addressed, as exemplified in Figure 3. No errors in interpretation of the statement, illogical associations, or hallucinations were identified. This result is in line with what is described in the literature, which shows that hallucinations are not as frequent in chatbots because they are designed to answer questions based on rules established during the programming phase and on the information contained in their databases, rather than to generate new information[15], which is usually the source of hallucinations. A similar study recently confirmed that tendency[16], which suggests that chatbots lack familiarity with the specificities and nuances of radiology, that lack of familiarity being the main obstacle to achieving higher accuracy rates.

The fact that ChatGPT performs better on questions that assess lower-order cognitive skills than on those that assess higher-order cognitive skills has been demonstrated in the literature[7] and was corroborated in the present study. This finding shows the ability of AI to recognize and express concepts and definitions while indicating that there are still advances to be made in terms of meeting more complex challenges. It is important that this characteristic of current AI models be known, so that future efforts can be directed toward increasing their performance in both orders of cognitive skills.

Large language models like ChatGPT are trained, from a large database, to recognize language patterns and the relationships between words. Therefore, the superior accuracy rate for physics questions over clinical questions observed in the present study is understandable. Because the ChatGPT database was not created specifically to meet the needs of radiologists, other areas of knowledge that transcend this specialty, such as physics, have the potential to generate a greater number of associations, thus increasing the accuracy rate for the challenges proposed. Such language models, including ChatGPT, could benefit from greater training in this medical specialty in the future. However, until then, it is important that radiologists be aware of this limitation.

Likewise, the absence of a statistically significant difference between radiology subspecialties can be understood as resulting from the limited familiarity that ChatGPT has with the terms and jargon employed in each of those areas. Radiology and each of its subspecialties have their own vernacular that is used in preparing reports, making classifications, and describing diagnoses. As long as the large language model database is not specifically trained to deal with these terms, the AI can be led to make incorrect associations, which limits its accuracy. For example, the word “density” has an obvious meaning for the radiologist, but it can be recognized by

ChatGPT as a different concept from that intended, simply because of the lack of training with the term in the specific context. Training in this specific technical language could improve the accuracy of AI, not only in radiology as a whole but also in its subspecialties.

Another noteworthy finding of the present study is the fact that ChatGPT analyzed all of the alternative answer choices for most of the questions presented. It is not clear what factor motivated the model to carry out such an analysis for some questions and not for others, given that the phenomenon was observed for questions related to all specialties and of all types, regardless of their characteristics. Nevertheless, when the analysis of the alternative answer choices is not done spontaneously, it is possible to ask ChatGPT in a subsequent message to carry out such an evaluation, and those requests were complied with 100% of the time in our study. This is a skill that can become useful for residents who wish to use the questions from previous tests, which are made available by the CBR, as study material. More than simply indicating the correct answer, the model tends to provide a complete study of the statements that make up the question, reviewing the topics covered in it, which indicates a possible role for ChatGPT as an auxiliary study tool, capable of succinctly yet efficiently reviewing topics of interest to radiology residents.

One of the differences between our findings and those of similar studies carried out in other countries is that the proportion of correct answers on questions related to the topic of physics was relatively high in our study. For example, ChatGPT provided the correct answer on 90% of the physics questions in our study, compared with only 40% in a study carried out in the United States[7]. Although it cannot be said with certainty, it is possible that the divergence is attributable to differences in the content of the questions (variations between the two countries in terms of the topics that are addressed within the field of physics) or in the process of their formulation (in this study, they were created by a specialized committee of the CBR, which is a national institution, whereas, in the study conducted in the United States case, the questions were created by researchers at a single center). In addition, although it is not yet clear, it is possible that the source language also has some influence on the performance of ChatGPT, given that there is greater availability of literature in English for training the model, which would therefore, theoretically, have less familiarity with questions in Portuguese. Furthermore, the translation performed by the model may not perfectly capture the meaning of some of the natural terms or expressions in Portuguese. As new studies in different languages appear, it is hoped that this topic will be elucidated.

This study has some limitations. Only objective, theoretical questions that did not involve the interpretation of radiological images were used, because ChatGPT does not yet have the capability to interpret images. The fact that that we provided feedback (correction) after each

error might have had an influence on the performance of ChatGPT; it is possible that its subsequent answers would have been different if there had been no such feedback. How much this interaction with the model affects the final result is a line of research that has yet to be explored. In addition, the number of questions related to each subspecialty was relatively small, which limits the comparison between these groups. Future studies with a greater number of questions could enrich this discussion.

## CONCLUSION

In summary, this study shows that, even without dedicated training in this area, ChatGPT presents reasonable performance, albeit still insufficient for approval, on radiology questions formulated by the CBR. It is expected that specific training in radiology for AI models such as ChatGPT will make their performance in matters of this specialty progressively better, and the radiology community must remain attentive to this evolution in order to take advantage of its potential.

**Acknowledgments**

The authors would like to thank the CBR Committee for Certification and Licensing, in the person of Dr. Tulio Augusto Alves Macedo, for authorizing the use of the questions formulated by the CBR.

## REFERENCES

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. medRxiv [Preprint]. 2023:2023.02.02.23285399.

5. O que é ChatGPT e por que alguns o veem como ameaça? Rio de Janeiro: G1; 2023. [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. Introducing ChatGPT. San Francisco: OpenAI; 2023. [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;0:0.

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? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023.

11. Weng TL, Wang YM, Chang S, et al. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023.

12. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. arXiv 2212.14882. [cited 2023 June 2]. Available from: https://arxiv.org/abs/2212.14882.

13. Avaliação anual de residentes – provas anteriores. São Paulo: Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, 2023 [cited 2023 Jun 2]. Available from: https://cbr.org.br/avaliacao-anual-de-residentes-provas-anteriores/.

14. Qasrawi R, BeniAndelrahman 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. Int Online J Educ Teach. 2020;7:744–58.

15. Alkaissi 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;0(0).

**Figure 1.** Example of a correct answer provided by ChatGPT on a question assessing a lower-order cognitive skill (a definition, in this case).

Select the option that represents the etiology of mosaic attenuation on a chest HRCT

A) Partial alveolar filling with interstitial thickening.
B) Thickening of bronchiolar walls with intraluminal mucus plugs.
C) Organizing pneumonia with pneumatoceles.
D) Bronchiolar obstruction with air trapping.
E) Thickening of the interlobular septa with intervening ground-glass opacities.

The correct option is E) Thickening of the interlobular septa with intervening ground-glass opacities.

Explanation: Mosaic attenuation is a radiological finding observed on high-resolution computed tomography (HRCT) of the chest. This appearance is characterized by the presence of areas of different lung attenuation, resulting in a mosaic pattern on the image.
The most common etiology of mosaic attenuation on chest HRCT is thickening of the interlobular septa with intervening ground-glass opacities. This alteration indicates the presence of inflammation and edema of the interstitial spaces of the lung, leading to differentiated attenuation of the affected areas.
The other options have different causes that are not directly related to mosaic attenuation:
A) Partial alveolar filling with interstitial thickening is most suggestive of pneumonia, pulmonary edema, or diffuse interstitial lung disease.
B) Thickening of the bronchiolar walls with intraluminal mucus plugs is more suggestive of bronchiectasis or bronchiolitis obliterans.
C) Organizing pneumonia with pneumatoceles is a condition characterized by pneumonia with the formation of air-filled cavities.
D) Bronchiolar obstruction with air trapping is more suggestive of bronchiolitis obliterans.

Therefore, the correct answer is option E) Thickening of the interlobular septa with intervening ground-glass opacities.

**Figure 2.** Example of a correct answer provided by ChatGPT on a question assessing a higher-order cognitive skill (the correlation between diseases, in this case).

Regarding laryngeal neoplasia, select the CORRECT option.

A) Supraglottic squamous cell carcinoma usually has a later clinical presentation and greater lymph node involvement than does glottic squamous cell carcinoma.
B) The larynx is the most common site for HPV-related squamous cell carcinoma to occur.
C) Invasion of the deep spaces affects the pre-epiglottic, paralaryngeal, and parapharyngeal spaces.
D) Invasion of the cricoid cartilage by laryngeal neoplasia classifies it as T3 in the TNM system.
E) The lymph nodes most affected in laryngeal neoplasia are those of levels I, II, and III.

The correct option is:
A) Supraglottic squamous cell carcinoma usually has a later clinical presentation and greater lymph node involvement than does glottic squamous cell carcinoma.

In laryngeal neoplasia, supraglottic squamous cell carcinoma generally presents symptoms later in comparison with glottic squamous cell carcinoma. This is because supraglottic lesions may initially go unnoticed, because symptoms such as hoarseness and dysphagia are less noticeable. Furthermore, supraglottic squamous cell carcinoma has a greater tendency to present lymph node dissemination than does glottic squamous cell carcinoma.
The remaining options are incorrect.
B) HPV-related squamous cell carcinoma is more common in other regions, such as the oropharynx, and it is not most common in the larynx.
C) In laryngeal neoplasia, invasion of deep spaces does not occur specifically in the spaces mentioned in this option.
D) The T3 classification in the TNM system refers to invasion of the larynx beyond the thyroid cartilage or invasion of the cricoid cartilage. Isolated invasion of the cricoid cartilage is classified as T2.
E) The lymph nodes most affected in laryngeal neoplasia are the cervical lymph nodes, which are grouped at different levels. Levels I, II, and III correspond to the lymph nodes of the neck.

**Figure 3.** Example of a ChatGPT error on an anatomy question in neuroradiology. The correct answer would be C.

Which cranial nerves pass through the superior orbital fissure?

A) I, II, III, and IV.
B) II, III, IV, and V1.
C) III, IV, V1, and VI.
D) IV, V1, V2, and VI.
E) IV, V1, V2, and V3.

The cranial nerves that pass through the superior orbital fissure are:
D) IV, V1, V2 and VI.

The superior orbital fissure is an opening located in the upper portion of the eye orbit. Through this fissure, the following cranial nerves pass:

- Trochlear nerve (IV): It is responsible for the movement of the superior oblique muscle of the eye.
- Ophthalmic branch of the trigeminal nerve (V1): It is responsible for sensitivity in the forehead region, upper eyelid, anterior part of the scalp, and cornea.
- Maxillary branch of the trigeminal nerve (V2): It is responsible for sensitivity in the cheek region, upper lip, part of the nose, upper jaw, and upper teeth.
- Abducens nerve (VI): It is responsible for the movement of the lateral rectus muscle of the eye.

Therefore, the correct option is D) IV, V1, V2, and VI.

**Figure 4.** Example of a correct answer provided by ChatGPT on a question in which there is a description of the imaging findings and a diagnosis is requested.

Consider a newborn, at 48 hours of age, with abdominal distension and no meconium elimination. On a simple X-ray of the abdomen, distension of several segments of intestinal loops is observed, with no air-fluid levels, no air in the rectum, and a mass effect in the direct lower quadrant with soft tissue density interspersed with small pockets of air. A barium enema was administered, showing that the colon had a mean caliber of 1.0 cm, with small filling defects. In this case, what is the likely diagnosis?

A) Meconium plug syndrome.
B) Meconium ileus.
C) Midgut volvulus.
D) Ladd's band.
E) Jejunal atresia.

In the case presented, the likely diagnosis is B) Meconium ileus.

Explanation:
Meconium plug syndrome (option A) usually presents as distal obstruction. with accumulation of meconium in the rectal ampulla. which was not described in the case.
Midgut volvulus (option C) usually occurs in older children and can cause complete intestinal obstruction, which was not mentioned in this case.
Ladd's band (option D) is a congenital anomaly related to malrotation of the intestine during fetal development and can lead to symptoms of obstruction but is not the most likely cause in the case presented.
Although jejunal atresia (option E) is also a possible cause of intestinal obstruction in newborns, the findings described in the case are not typical of that condition.
Meconium ileus (option B) is a complication associated with cystic fibrosis, in which the meconium is abnormally thick and obstructs the bowel loops. The presence of abdominal distension, absence of meconium elimination, distension of segments of intestinal loops on X-ray, narrowed colon on barium enema and soft tissue density interspersed with small pockets of air are characteristics consistent with meconium ileus.
Therefore, based on the information provided, the likely diagnosis is meconium ileus. However, it is important to highlight that the definitive diagnosis and treatment must be carried out by a specialist after a complete clinical evaluation and appropriate complementary examinations.

**Figure 5.** Example of a correct answer provided by ChatGPT on a question in which there is a description of a clinical case with imaging examination and the most appropriate course of action is requested.

Analyze the clinical case below.

A 20-year-old male patient reports routine steroid use to increase muscle mass and complains of increased bilateral breast volume. The mammogram shows only fibroglandular densities in both retroareolar regions and the ultrasound reveals a solid, hypoechoic. heterogeneous, palpable nodule, measuring 0.6 cm, in the subareolar region of the right breast. Considering this information, select the correct option.

A) This is a case of gynecomastia, and the patient should be advised to stop using steroids.
B) Follow-up of the nodule within 6 months and reevaluation of the gynecomastia with mammography should be recommended after a reduction in breast volume is achieved through interruption of the steroid use.
C) Determination of the evolution of the nodule in the right breast after 6 months should be recommended, with no need for a new mammogram.
D) A biopsy of the nodule in the right breast should be recommended.
E) The next examination to be performed is magnetic resonance imaging.

Considering the information presented, the correct option is A) This is a case of gynecomastia, and the patient should be advised to stop using steroids.
Explanation: The patient is a 20-year-old man who reports using steroids to increase muscle mass and presents an increase in bilateral breast volume. Mammography shows only fibroglandular densities in the retroareolar regions, and ultrasound reveals a solid, hypoechoic, heterogeneous, palpable nodule, measuring 0.6 cm, in the subareolar region of the right breast.
These findings are consistent with gynecomastia, which is an increase in breast tissue in men due to a hormonal imbalance. The use of anabolic steroids is one of the known causes of gynecomastia in young men who practice bodybuilding. The presence of a solid nodule in the subareolar region is common in gynecomastia, and mammography usually shows only fibroglandular densities, without suspicious findings.
Therefore, in this case, the appropriate approach will be to advise the patient to stop using steroids, given that they may be contributing to the development of gynecomastia. It is not necessary to perform a biopsy of the nodule, because the characteristics described are consistent with gynecomastia and do not suggest malignancy. Monitoring may be recommended to assess the regression of gynecomastia after the steroid use has been discontinued, but it is not necessary to perform a new mammogram. Magnetic resonance imaging is not routinely indicated in the diagnosis of gynecomastia.

**Table 1**—ChatGPT performance by question type and topic.

| Question characteristic | Questions n | Correct answers n (%) | *p* |
|---|---|---|---|
| Type | | | |
| Lower-order cognitive skills | 59 | 38 (64.4) | |
| Higher-order cognitive skills | 106 | 50 (47.2) | |
| Description of findings | 42 | 22 (52.4) | 0.81* |
| Clinical management | 22 | 12 (54.5) | 0.72* |
| Application of a concept | 57 | 38 (66.7) | 0.67* |
| Calculation/classification of findings | 8 | 3 (37.5) | 0.92* |
| Correlations between diseases | 26 | 11 (42.3) | 0.63* |
| Anatomy | 10 | 2 (20.0) | 0.58* |
| Topic | | | 0.02* |
| Physical | 20 | 18 (90.0) | |
| Clinical | 145 | 68 (46.8) | 0.41* |
| Abdominal imaging | 20 | 13 (65.0) | 0.62† |
| Thoracic imaging | 15 | 9 (60.0) | 0.56† |
| Neuroradiology | 15 | 5 (33.3) | 0.76† |
| Musculoskeletal imaging | 15 | 8 (53.3) | 0.87† |
| Breast imaging | 15 | 7 (46.7) | 0.61† |
| Contrast media | 15 | 9 (60.0) | 0.94† |
| Ultrasound | 15 | 3 (20.0) | 0.78† |
| Pediatric radiology | 15 | 10 (66.7) | 0.93† |
| Obstetrics and gynecological imaging | 10 | 2 (20.0) | 0.72† |
| Miscellaneous | 10 | 4 (40.0) | 0.65† |
| Total | 165 | 88 (53.3) | 0.01* |

*Fisher's exact test.

†Analysis of variance.

# APÊNDICE 4 - SEGUNDO ARTIGO

**Performance of ChatGPT on residency questions: insights into artificial intelligence and medical education**

**ABSTRACT**

**Introduction:** Recent advances in artificial intelligence raise questions about its use in medicine, but its potential to assist in medical education remains uncertain. ChatGPT, a publicly available online chatbot, represents one such advancement. Understanding its familiarity with healthcare topics can help define its potential future use not only in medicine but also in medical education.

**Objective:** To test the accuracy of ChatGPT in solving medical residency exam questions and analyze, based on this performance, its implications for medical education.

**Method:** 500 medical residency exam questions from a national hospital (2017-2021) were presented to ChatGPT, and its responses were analyzed. Questions were divided, for statistical analysis, into those assessing lower or higher-order cognitive skills and according to specialty and question type. Student's t-tests, chi-square tests, and ANOVA were conducted.

**Result:** ChatGPT correctly answered 56% of the questions (280 out of 500). There was a statistically significant difference in performance between questions assessing lower-order cognitive skills (63.8%, 99 out of 155) and higher-order skills (52.5%, 181 out of 345) (P= 0.01). The best performance was in internal medicine (70%), followed by pediatrics (57%), general surgery (56%), gynecology and obstetrics (51%), and preventive and social medicine (46%), with no significant difference between these areas or among different question types (P > 0.01).

**Discussion:** ChatGPT's performance exceeded the passing threshold in only four of the exam specialties. The displayed knowledge suggests that students could use ChatGPT to review various topics, especially because it tends to conduct a thorough analysis of all options in each question. It also exhibits satisfactory synthesis capabilities, summarizing information from consensuses and guidelines, making it easier for students to access this information.

**Conclusion:** ChatGPT demonstrated performance below that of most actual candidates in medical residency exam questions. The analysis of its responses and abilities suggests a potential supportive role for this tool in assisting medical students in studying the covered subjects.

**KEYWORDS: Artificial intelligence; medical education; exam questions.**

**INTRODUCTION:**

"Artificial Intelligence" (AI) is the general term given to computing methods that simulate the learning patterns of human intellect[1]. The recent rapid advancements in this field have raised questions about how it will impact various professions in the future, including medicine. Among existing artificial intelligence products, ChatGPT (Chat Generative Pre-trained Transformer) has gained prominence not only in scientific literature[2,3,4] but also in mainstream media[5]. It is an artificial intelligence tool based on relationships between AI algorithms and human language, known as Natural Language Processing (NLP), publicly available since November 30, 2022[6]. Its currently available algorithm is GPT-3.5, a large language model trained with over 45 terabytes of textual data. Through neural networks, these data enable the tool to analyze and generate texts similar to those written by humans[7]. Although not specifically trained for use in healthcare, studies have already demonstrated its promising role in both medical practice[8] and academic writing in medicine[9]. To assess ChatGPT's familiarity with healthcare topics, the tool has been tested in academic exams evaluating real students, such as the United States Medical Licensing Examination[10] and questions for obtaining specialist titles in radiology in Canada/United States[7] and family medicine in Taiwan[11], with results showing overall performance close to the required passing level.

Despite the numerous recent studies on the use of ChatGPT in medicine, there are still few papers regarding its role in medical education, and none with national data. The tool's ability to engage with a student as if it were another student, simulate a patient, and provide feedback for received questions represent potential ways to make it useful in the training of new physicians, provided its familiarity with the topics being studied is adequate[10]. This study aims to evaluate these issues in the Brazilian context by testing ChatGPT's performance on topics frequently explored in exams for admission to a national medical residency program, analyzing its responses, and how they can be used to assist in the education of medical students.

**OBJECTIVE:**

To test the accuracy of a widely available artificial intelligence platform (ChatGPT) in solving questions applied in exams for admission to a medical residency program and analyze the potential of this tool in supporting medical education.

**METHOD:**

Prospective analytical study conducted between May 31 and June 10, 2023, which did not involve humans or patient data, and institutional ethics committee approval was waived.

**Medical residency questions**

500 questions from exams for admission to a national hospital's medical residency program conducted between 2017 and 2021 were selected. These questions are publicly available online on the website of the Exam Center of this institution[12]. All questions were of the "multiple-choice" type, with only one correct answer and four false alternatives. To ensure representation of all major medical areas, 100 questions were chosen from each of the following: internal medicine, general surgery, pediatrics, gynecology and obstetrics, and preventive and social medicine. Subsequently, the questions were subdivided according to Bloom's taxonomy[13] principles into questions assessing lower-order cognitive skills (remembering an idea, memorizing a concept) and questions assessing higher-order cognitive skills (evaluating, analyzing, synthesizing knowledge). These were further divided according to their style into seven subcategories: (1) description of a clinical finding or sign, (2) clinical management of a patient, (3) application of a concept, (4) calculation or classification of described findings, (5) physiopathological review of one or more diseases, (6) indicating the specific diagnosis, and (7) indicating the specific treatment. All questions were independently classified by the study authors, and in cases of disagreement, a final classification was obtained through consensus.

**ChatGPT**

The GPT-3.5 version of ChatGPT (May 24, 2023; OpenAI) was used, since it was the most advanced version avalaible at the time of this study. Although this tool was trained with over 45 terabytes of text data from internet pages, books, and scientific articles, it was not specifically provided for medical needs. ChatGPT does not perform internet searches, responding to questions using only its existing database.

**Data colection and analysis**

The questions and their respective alternatives were presented to ChatGPT one by one, and its responses were recorded for subsequent analysis by researchers. For questions answered incorrectly, immediate feedback was provided, explaining the error and the correct answer, to analyze the platform's behavior. In addition to the quantitative analysis of the number of correct and incorrect answers, researchers conducted a qualitative group analysis, reaching a consensus on comments regarding the obtained responses.

**Statistical analysis**

For accuracy analysis, the ratio of the number of correct answers to the total number of questions for all categories (total questions, high and low-order questions, and question subtypes as described above) was calculated, and the final percentage of this ratio was displayed.

Statistical analysis used t-tests, chi-square, and ANOVA tests to correlate question groups with each other and different categories separately. The STATA 16 program was used, and post-processing

was carried out in Microsoft Excel 365, with its data analysis package. Values of $p < 0.05$ were considered statistically significant.

## RESULTS

### Overall performance

ChatGPT had an overall accuracy of 56% (correctly answered 280 out of 500 presented questions). Considering the five-year average, this score would allow a residency candidate to progress to the second phase of the competition in only four of the sixteen medical specialties offered. TABLE 1 compares the platform's score with the average cutoff score achieved by candidates in the same five competitions for different specialties.

**Table 1.** Comparison between ChatGPT performance and the cutoff score for advancing to the next phase in the medical residency competition in different medical specialties.

| Specialty | Avarage cutoff score (2017-2021) |
|---|---|
| Dermatology | 71 |
| Ophthalmology | 68,6 |
| Internal medicine | 68 |
| Otorhinolaryngology | 68 |
| Neurosurgery | 67,3 |
| Anesthesiology | 65,6 |
| General surgery | 66,3 |
| Psychiatry | 64,3 |
| Neurology | 64 |
| Pediatrics | 62 |
| Gynecology/Obstetrics | 61 |
| Radiology | 60,3 |
| **ChatGPT** | **56** |
| Infectious Diseases | 53,6 |
| Orthopedics | 51,3 |
| Parthology | 49 |
| Family medicine | 40 |

Source: Compiled by the authors.

### Performance by Question Type

The accuracy in questions assessing lower-order cognitive skills (63.8%, 99 out of 155) was higher than that observed in questions evaluating higher-order cognitive skills (52.5%, 181 out of 345) ($p = 0.01$).

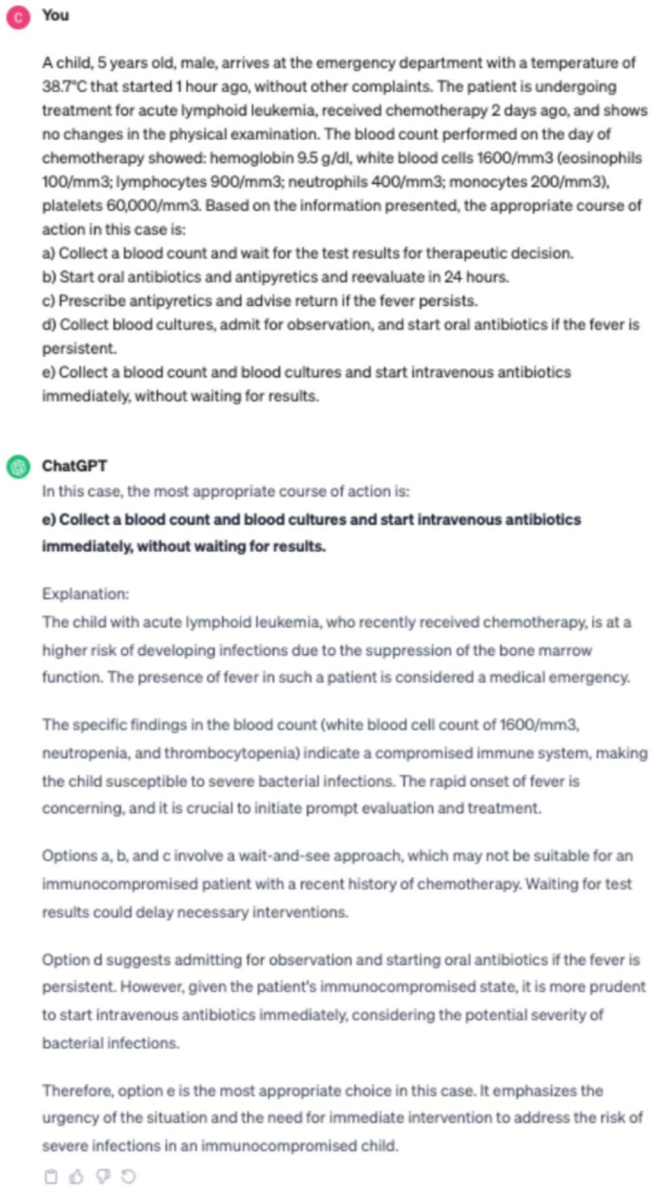
The performance according to the question style is summarized in TABLE 2, with no statistically significant difference observed among various styles ($p > 0.05$). FIGURES 1 and 2 illustrate examples of ChatGPT's correct responses to these different types of questions.

**TABLE 2.** ChatGPT Performance According to Question Style.

| Question Type | Total N | Correct Answers | Accuracy (%) |
|---|---|---|---|
| Description | 19 | 14 | 74 |
| Clinical Management | 121 | 65 | 54 |
| Application of Concept | 133 | 69 | 52 |
| Calculation or Classification | 26 | 16 | 61 |
| Theoretical Review of the Topic | 116 | 57 | 49 |
| Indicate Diagnosis | 56 | 38 | 68 |
| Indicate Treatment | 29 | 21 | 72 |

Source: Compiled by the authors.

**FIGURE 1.** Example of a higher-order question in the "clinical management" style (indicate the correct approach to a clinical scenario).



**You**

A child, 5 years old, male, arrives at the emergency department with a temperature of 38.7°C that started 1 hour ago, without other complaints. The patient is undergoing treatment for acute lymphoid leukemia, received chemotherapy 2 days ago, and shows no changes in the physical examination. The blood count performed on the day of chemotherapy showed: hemoglobin 9.5 g/dl, white blood cells 1600/mm3 (eosinophils 100/mm3; lymphocytes 900/mm3; neutrophils 400/mm3; monocytes 200/mm3), platelets 60,000/mm3. Based on the information presented, the appropriate course of action in this case is:
a) Collect a blood count and wait for the test results for therapeutic decision.
b) Start oral antibiotics and antipyretics and reevaluate in 24 hours.
c) Prescribe antipyretics and advise return if the fever persists.
d) Collect blood cultures, admit for observation, and start oral antibiotics if the fever is persistent.
e) Collect a blood count and blood cultures and start intravenous antibiotics immediately, without waiting for results.

**ChatGPT**

In this case, the most appropriate course of action is:

**e) Collect a blood count and blood cultures and start intravenous antibiotics immediately, without waiting for results.**

Explanation:
The child with acute lymphoid leukemia, who recently received chemotherapy, is at a higher risk of developing infections due to the suppression of the bone marrow function. The presence of fever in such a patient is considered a medical emergency.

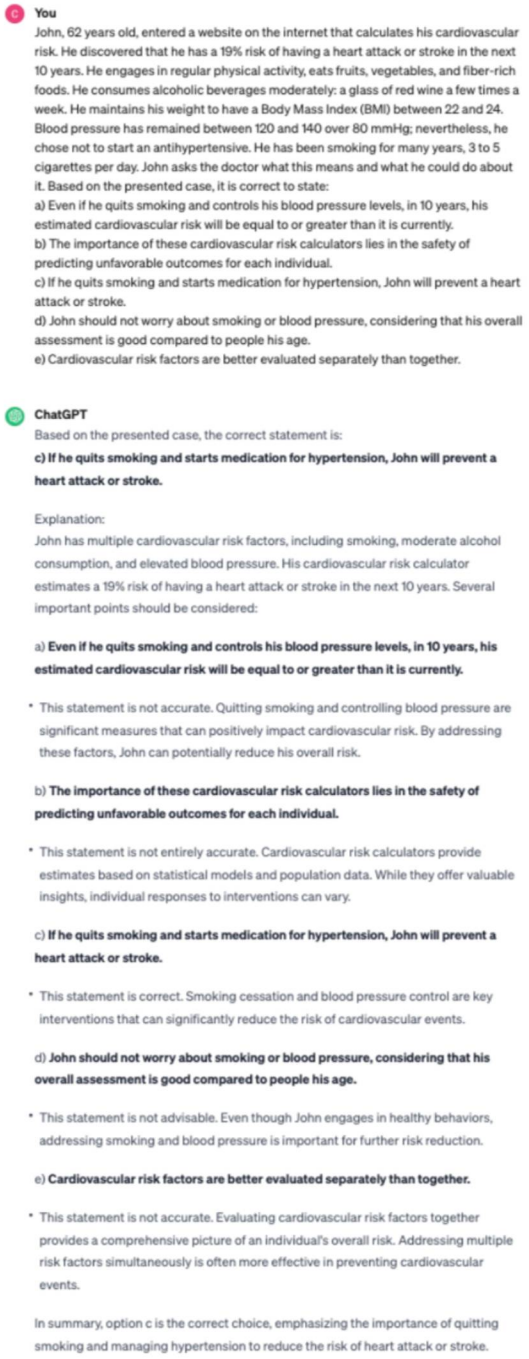
The specific findings in the blood count (white blood cell count of 1600/mm3, neutropenia, and thrombocytopenia) indicate a compromised immune system, making the child susceptible to severe bacterial infections. The rapid onset of fever is concerning, and it is crucial to initiate prompt evaluation and treatment.

Options a, b, and c involve a wait-and-see approach, which may not be suitable for an immunocompromised patient with a recent history of chemotherapy. Waiting for test results could delay necessary interventions.

Option d suggests admitting for observation and starting oral antibiotics if the fever is persistent. However, given the patient's immunocompromised state, it is more prudent to start intravenous antibiotics immediately, considering the potential severity of bacterial infections.

Therefore, option e is the most appropriate choice in this case. It emphasizes the urgency of the situation and the need for immediate intervention to address the risk of severe infections in an immunocompromised child.

**FIGURE 2.** Example of a higher-order question in the "calculation or classification" style (interpret the classification of cardiovascular risk). In this case, the correct answer would be letter "A" (since aging is a factor that increases cardiovascular risk independently).



**You**

John, 62 years old, entered a website on the internet that calculates his cardiovascular risk. He discovered that he has a 19% risk of having a heart attack or stroke in the next 10 years. He engages in regular physical activity, eats fruits, vegetables, and fiber-rich foods. He consumes alcoholic beverages moderately: a glass of red wine a few times a week. He maintains his weight to have a Body Mass Index (BMI) between 22 and 24. Blood pressure has remained between 120 and 140 over 80 mmHg; nevertheless, he chose not to start an antihypertensive. He has been smoking for many years, 3 to 5 cigarettes per day. John asks the doctor what this means and what he could do about it. Based on the presented case, it is correct to state:
a) Even if he quits smoking and controls his blood pressure levels, in 10 years, his estimated cardiovascular risk will be equal to or greater than it is currently.
b) The importance of these cardiovascular risk calculators lies in the safety of predicting unfavorable outcomes for each individual.
c) If he quits smoking and starts medication for hypertension, John will prevent a heart attack or stroke.
d) John should not worry about smoking or blood pressure, considering that his overall assessment is good compared to people his age.
e) Cardiovascular risk factors are better evaluated separately than together.

**ChatGPT**

Based on the presented case, the correct statement is:

**c) If he quits smoking and starts medication for hypertension, John will prevent a heart attack or stroke.**

Explanation:
John has multiple cardiovascular risk factors, including smoking, moderate alcohol consumption, and elevated blood pressure. His cardiovascular risk calculator estimates a 19% risk of having a heart attack or stroke in the next 10 years. Several important points should be considered:

a) **Even if he quits smoking and controls his blood pressure levels, in 10 years, his estimated cardiovascular risk will be equal to or greater than it is currently.**

- This statement is not accurate. Quitting smoking and controlling blood pressure are significant measures that can positively impact cardiovascular risk. By addressing these factors, John can potentially reduce his overall risk.

b) **The importance of these cardiovascular risk calculators lies in the safety of predicting unfavorable outcomes for each individual.**

- This statement is not entirely accurate. Cardiovascular risk calculators provide estimates based on statistical models and population data. While they offer valuable insights, individual responses to interventions can vary.

c) **If he quits smoking and starts medication for hypertension, John will prevent a heart attack or stroke.**

- This statement is correct. Smoking cessation and blood pressure control are key interventions that can significantly reduce the risk of cardiovascular events.

d) **John should not worry about smoking or blood pressure, considering that his overall assessment is good compared to people his age.**

- This statement is not advisable. Even though John engages in healthy behaviors, addressing smoking and blood pressure is important for further risk reduction.

e) **Cardiovascular risk factors are better evaluated separately than together.**

- This statement is not accurate. Evaluating cardiovascular risk factors together provides a comprehensive picture of an individual's overall risk. Addressing multiple risk factors simultaneously is often more effective in preventing cardiovascular events.

In summary, option c is the correct choice, emphasizing the importance of quitting smoking and managing hypertension to reduce the risk of heart attack or stroke.

**Performance by Medical Specialty**

ChatGPT's highest accuracy was achieved in internal medicine questions (70%), followed by pediatrics (57%), general surgery (56%), gynecology and obstetrics (51%), and preventive and social medicine (46%). Statistical analysis did not reveal a significant difference between these areas ($p > 0.05$).

**Qualitative Evaluation of Responses**

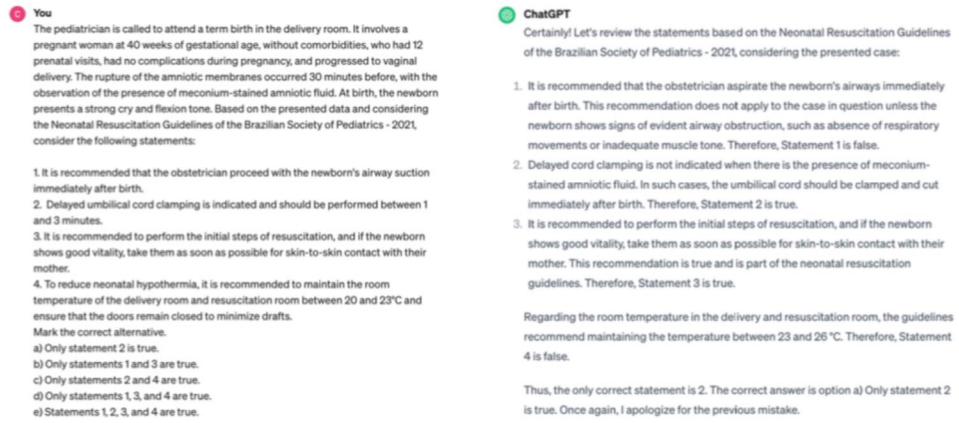
The unanimous assessment by evaluators was that ChatGPT's performance was satisfactory, especially considering that its database was not specifically developed for use by healthcare professionals. The high degree of assertiveness in providing answers, never using words indicating possible doubt or hesitation (FIGURES 1 and 2), even in questions where an incorrect answer was marked (FIGURE 2), was noteworthy. Another interesting aspect is that in the majority of questions (67%, 338 out of 500), the program not only indicated the correct answer but also analyzed all options, explaining why it considered them incorrect (FIGURES 1 and 2).

**DISCUSSION**

In this study, the first conducted with exclusively Brazilian data, it is evident that ChatGPT demonstrates reasonable accuracy but is insufficient for approval in most medical residency programs. The better performance in questions assessing lower-order cognitive skills compared to higher-order ones has been previously demonstrated in the literature[7], showcasing AI's ability to recognize and express concepts and definitions, with further advancements needed in more complex challenges. New training and refinement of its database may lead to improved performance in both types of cognitive order. The observed statistical difference in performance among various question types not only indicates that correct answers did not occur by chance but also reinforces the versatility of artificial intelligence in addressing different problem types (deciding on approaches, indicating treatments, correlating diseases, etc.). Similarly, the absence of a significant difference in performance among different medical specialties indicates that ChatGPT does not exhibit any specific inclination or limitation towards any of these domains. The errors presented by ChatGPT can be explained by at least two aspects. The first is the presence of phrases and terms (referred to in computer language as "lemmas") in the questions that are similar or even redundant to lemmas that activate another pathway of association in the algorithm. An example is illustrated in FIGURE 2, where the phrase "will prevent a new heart attack" made sense to the algorithm, even though it is not the correct answer, as the lifestyle changes present in the option activate association pathways leading to the conclusion that there will be prevention of cardiovascular events. Another factor related to the discrepancy between real performance and the previously established performance of software like ChatGPT is related to a mathematical analysis mechanism called Overfitting[18]. Overfitting occurs when the model is trained and performs well with its training set (which was

generated), but struggles to generalize in real-world applications, such as residency questions. This occurs in complex models like ChatGPT, or due to prolonged training periods, leading the algorithm to "memorize" training data but struggle to apply it correctly in real scenarios. The performance obtained in national questions was similar to similar studies conducted abroad[10,11,14]. These results can be considered encouraging, as they show that there is still room for AI to be trained in health sciences, and once enriched in this area, programs will likely achieve even better performances. The impact of this new reality on the practice of medicine is still under discussion[15,16], pointing out the main evidence for its potential diagnostic, therapeutic, and health surveillance uses[15]. Its role in medical education, however, seems particularly promising. For example, the way the platform responded to questions shows that it is possible for a student to use it for academic content review. In this study, ChatGPT often not only marked the correct alternative in the questions presented but also conducted a detailed analysis of all options in the question (67%, 338 out of 500) (FIGURES 1 and 2). This type of analysis offers students a more detailed review of the topic than a simple answer about which alternative is correct, which can help them better understand and retain the material covered in the exercise. ChatGPT also has satisfactory summarization power, capable of summarizing information available in consensus and guidelines established by expert societies. After being fed articles and consensuses on a specific topic, it is possible to ask artificial intelligence for a synthesis of the information contained in them quickly and efficiently[17], dispensing with searching and reviewing all files individually, which would take more time for the student. An example of this ability was seen in this study, where ChatGPT was able to cite the national guidelines that were being referenced by the question[19] (FIGURE 3). However, in the same question, the platform contradicted itself, judging two options as true during analysis but marking only one as correct when answering the question, demonstrating that its operation is not perfect.

**FIGURA 3.** Example of ChatGPT referencing a guideline.




This study has limitations. Only questions from a single institution were used, which may limit the topics covered by the questions, according to what the exam board values the most. New studies with questions from other regions of the country could be useful to better test the limits of ChatGPT's database. Additionally, only objective questions were included, which make up only one phase of the residency admission competition, excluding essay questions and practical stations that also usually compose such exams. The qualitative evaluation of questions conducted in this study has a high degree of subjectivity, and other researchers may have different impressions, opening the way for further research on the subject.

## CONCLUSIONS

In summary, ChatGPT showed performance comparable to real candidates in national medical residency competition questions, but still insufficient for approval in most programs. The analysis of its responses suggests a potential auxiliary role for this tool in helping medical students study the topics covered. The academic community should remain vigilant about ChatGPT and other large language models to harness their potential capabilities.

## REFERENCES

1. Wang F, Preininger A. AI in health: state of the art, challenges, and future directions. Yearbook of medical informatics. 2019;28:16–26.
2. Morreel S, Mathysen D, Verhoeven V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. Med Teach. 2023 Jun;45(6):665-666.
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, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv [Preprint]. 2023 Feb 7:2023.02.02.23285399.

5. O que é ChatGPT e por que alguns o veem como ameaça? [acesso em 10 jun 2023]. Disponível em: https://g1.globo.com/tecnologia/noticia/2023/01/19/o-que-e-chatgpt-e-por-que-alguns-o-veem-como-ameaca.ghtml.

6.Introducing ChatGPT. OpenAI. [acesso em 02 jun 2023]. Disponível em: 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;0:0.

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 May 4;6:1169595.

9. Biswas S. ChatGPT and the Future of Medical Writing. Radiology. 2023 Apr;307(2):e223312. doi: 10.1148/radiol.223312.

10. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023.

11. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023 Jun 9.

12. Núcleo de Concursos [internet]; c2017-2021 [acesso em 31 mai 2023]. Disponível em: https://rb.gy/1nr01.

13. Qasrawi R, BeniAndelrahman 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 (IOJET). 2020;7(3). 744-758.

14. Ali R, Tang OY, Connolly ID, Sullivan PLZ, Shin ZH, Fridley JS et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. medRxiv 2023.03.25.23287743; doi: https://doi.org/10.1101/2023.03.25.23287743.

15. King MR. The Future of AI in Medicine: A Perspective from a Chatbot. Ann Biomed Eng. 2023;51, 291–295.

16. Valente M, Bellini V, Del Rio P, Freyrie A, Bignami E. Artificial Intelligence Is the Future of Surgical Departments…. Are We Ready? Angiology. 2023;74(4):397-398.

17. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. Resuscitation. 2023;185:109729.

18. Tillmann AM. On the Computational Intractability of Exact and Approximate Dictionary Learning». IEEE Signal Processing Letters. 2023;22 (1):45-49

19. Sociedade Brasileira de Pediatria. Programa de Reanimação Neonatal [internet]. Recomendações para assistência ao recém-nascido na sala de parto [acesso em 04 jun 2023]. Disponível em: www.sbp.com.br/reanimacao.