ACM Arquivos Catarinenses de Medicina
ISSN (impresso) 0004-2773
ISSN (online) 1806-4280
AMB Associação Médica Brasileira

RESUMOS PREMIADOS NO CONGRESSO CATARINENSE DAS LIGAS ACADÊMICAS

DESEMPENHO DO CHATGPT 3.5 PARA RESPONDER QUESTÕES DE MÚLTIPLA-ESCOLHA EM ORTOPEDIA E TRAUMATOLOGIA

PERFORMANCE OF CHATGPT 3.5 TO ANSWER MULTIPLE-CHOICE QUESTIONS IN ORTHOPEDICS AND TRAUMATOLOGY

Vitor Ricardo Thais Malvezzi
Rebecca Fantuzzi
Sarah Domitila Menezes Mourão
Jean Canhete
Franciele Cascaes da Silva

RESUMO

Introdução: A IA e o ChatGPT oferecem inúmeras possibilidades que envolvem ensino médico, prática clínica e cirúrgica. A adição de agentes com capacidade de análise lógica amplia ainda mais essas possibilidades. No entanto, a IA também é suscetível a falhas e erros. Já o desempenho do GPT-3.5 em responder questões ortopédicas é desconhecido. **Objetivo:** Analisar o desempenho do GTP-3.5 para responder questões de múltiplas escolhas em Ortopedia e Traumatologia. Método: Foram utilizadas questões de ortopedia múltipla-escolha aplicadas em provas de uma universidade de medicina no sul do Brasil. Após seleção, restaram 66 questões. Essas foram classificadas subjetivamente por conteúdo e objetivamente por tamanho, seja de seu enunciado e seja pelo comprimento das alternativas. Finalmente, foi avaliada a veracidade da alternativa escolhida, classificando-as como corretas ou incorretas. **Resultados:** Dentro deste estudo observou-se uma média total de assertividade de 43,93% das questões, enquanto que 56,06% foram respondidas incorretamente. O grupo de questões com enunciados longos com alternativas longas, destacou-se positivamente com 54,55% de acertos, e dentre eles as questões sobre conduta obtiveram 62,5% de acertos. Por outro lado, enunciados curtos e alternativas longas apresentaram 68,75% de erros. **Conclusão:** A união dos modelos chatbot de inteligência artificial como o GPT-3.5 e a prática médica ortopédica/acadêmica aparenta não estar pronta. Por mais que a média percentual de respostas corretas em determinados conjuntos ultrapasse 50%, em geral não se observou tal resultado. Os ortopedistas, residentes e acadêmicos de medicina devem estar cientes da limitação que hoje possui o GPT-3.5.

Descritores: IA,ChatGPT, Medicina, Ortopedia.

INTRODUÇÃO

A criação do computador ampliou a capacidade do homem de calcular e armazenar inúmeras informações em nanosegundos (Da Silva, 2019)[1]. Os primórdios datam 1950, quando as ideias de Alan Turing em, reconhecido pelo seu trabalho na quebra dos códigos nazistas durante a segunda guerra mundial, levou os primeiros estudos no ramo da ciência da computação voltados para Inteligência

artificial (IA), que propõe desenvolver sistemas capazes de simular a percepção humana, sendo possível reconhecer o problema, seus componentes e os solucionar com propostas e tomadas de decisões (LOBO, 2018)[2].

A Inteligência Artificial Generativa (IA Generativa), ramo da IA, é capaz de criar diversos conteúdos e foi criada com intuito de fornecer resultados construtivos e coerentes de forma rápida e prática, a partir de combinações de palavras com maior probabilidade de estarem corretas com base em uma "entrada", pergunta ou enunciado (Ramos, 2023)[3]. Esses modelos de IA Generativa estão sendo utilizados na área médica para geração de relatórios, suporte educacional, suporte à decisão clínica, comunicação e análise de dados, como por exemplo, o Transformador Generativo Pré-treinado em bate-papo - ChatGPT (Tang, 2021)[4]. O ChatGPT apresenta grande potencial para melhorar lacunas de conhecimento presente em interações humanas, incluindo o âmbito da prática médica, tendo em vista que fornece respostas ajustadas para compreensão utilizando linguagem natural e convencional.[4]

De acordo com o artigo de Tang, Yang, Shajudeen, et al. (2021)[4] que comparou resultados do ChatGPT com as respostas de educação padrão referência sobre radiologia e impressão 3D médica, esse programa teve maior precisão do que outros modelos de IA, no entanto nenhum deles apresentou eficiência máxima. E, mesmo com respostas semelhantes a humanas, não houve respostas a todas as perguntas de maneira inteiramente correta e concluíram que o ChatGPT apresentou maior precisão quando feitas perguntas médicas mais simples e piorou, significativamente, seus resultados de acordo com o aumento da complexidade das perguntas, mostrando-se eficaz quando utilizado como apoio secundário, passando pela verificação e avaliação de um profissional de saúde qualificado na área.

A IA envolve diversas competências que são capazes de reconhecer padrões e imagens, perceber relações e nexos, seguir algoritmos de decisão propostos por especialistas, sendo algo além do processamento de dados. Assim, IA foi vista como ferramenta de possível potencial para atuar em serviço de triagem, dando suporte e auxílio nas decisões dos radiologistas, uma vez que a dimensão da incidência de fraturas na população como um todo, o diagnóstico tardio ou perdido de fraturas nas radiografias é um erro comum que varia de 3% a 10% (Kuo, 2022)[5].

Kuo RYL, et al. (2022)[5], revisaram e analisaram 42 estudos para comparar o desempenho do diagnóstico de fraturas pela IA e por médicos em publicações revisadas em radiografias e tomografias computadorizadas. Seus resultados demonstraram que a Inteligência Artificial, com 91% de sensibilidade e especificidade, apresentou alta precisão diagnóstica, tendo desempenho comparável ao dos médicos. Além disso, sua utilização associada agregou para a precisão e rapidez diagnóstica dos profissionais. Dessa maneira, a IA aproxima-se, em termos de desempenho de detecção de fraturas, aos médicos e ressalta, assim, que a IA atual deve ser utilizada como um complemento de diagnóstico, porém, não substitui a experiência prática e clínica dos profissionais.Portanto, é necessário equilibrar a

inovação com a necessidade de evidências a longo prazo e regulamentação adequada, e a educação e a certificação também desempenharam um papel importante no uso adequado dessas ferramentas digitais na medicina esportiva (Rigamonti,2020)[6].

Assim, para auxiliar nas evidências sobre a segurança do uso do ChatGPT na prática médica ortopédica, o estudo teve como objetivo analisar o desempenho do GTP-3.5 para responder questões de múltiplas escolhas em Ortopedia e Traumatologia.

MÉTODO

Trata-se de um estudo observacional transversal realizado a partir de questões de múltipla escolha "a b c d e" aplicadas em provas de ortopedia no curso de graduação em Medicina em uma universidade da região sul do Brasil no ano de 2023. Realizou-se a seleção das perguntas por meio da inclusão das questões de provas disponibilizadas, e exclusão das questões que continham imagem ou que haviam sido anuladas devido a ausência de um único gabarito correto.

Posteriormente, enviaram-se as perguntas para o GPT-3.5 por meio de dois pesquisadores e diferentes dispositivos, em novembro de 2023.

Tendo-se como base as questões enviadas, o GPT-3.5 forneceu uma escolha dentre as alternativas: quando igual a do gabarito, considerou-se correta, quando diferente, incorreta.

Quando o GPT3.5 optou por duas ou mais alternativas para a mesma questão considerou-se "resposta incorreta", mesmo que a resposta correta estivesse inclusa nessas 2 alternativas escolhidas.

O presente estudo classificou as questões em grupos distintos: referente ao conteúdo da questão (conhecimentos gerais, raciocínio diagnóstico ou tratamento/conduta), quanto ao tamanho do enunciado (curto quando igual ou menor que 78 caracteres e longo quando maior que 78 caracteres) e em relação ao tamanho das alternativas propostas (curtas, quando todas continham um número de caracteres igual ou menor a 83 e longa, quando uma ou mais alternativas ultrapassassem 83 caracteres).

RESULTADOS

Neste estudo, foi realizado um total de 66 questionamentos à plataforma de inteligência artificial ChatGPT, GPT-3.5. A partir dos resultados obtidos, destaca-se, que a porcentagem de acertos das questões foi de 43,93% (n=29 questões). Enquanto que 56,06% (n=37 questões) tiveram resposta divergente do estipulado pelo gabarito oficial, sendo consideradas incorretas.

Além disso, tornou-se importante identificar e destacar as respostas duplicadas obtidas, que ocorreram em 5,97% (n=4) do total de questões. Em cada uma dessas, o GPT-3.5 atribuiu a veracidade corretamente a uma afirmação e, erroneamente, à outra, resultando em uma resposta final considerada incorreta.

Notou-se, também, que diante da divisão das questões em 3 categorias de conteúdo a que obteve maior porcentagem de erro pela plataforma foi a de “conhecimentos gerais”, com 72,41% (n=21) questões respondidas incorretamente. Essa, seguida pela categoria de “conduta”, com 45,45% (n=5) e pela de “raciocínio diagnóstico”, com 42,30% (n=11)(Tabela 4).

A análise das porcentagens de acerto e erro das questões categorizadas de acordo com o tamanho do enunciado e das alternativas, permitiu observar resultados distintos. Nas questões caracterizadas por enunciados curtos e alternativas curtas, a taxa de acerto foi de 40% (n=4), enquanto a taxa de erro foi de 60% (n=6). Por outro lado, em questões com enunciados curtos e alternativas longas, 31,25% (n=5) das questões foram respondidas corretamente, enquanto os outros 68,75% (n=11) não foram. No contexto de enunciados longos e alternativas curtas, as porcentagens de acerto e erro ficaram em 46,67% (n=14) e 53,33% (n=16), respectivamente. Por fim, a maior taxa de acerto foi observada na categoria de enunciados longos com alternativas longas, com 54,55% (n=6), acompanhada de uma taxa de erro de 45,45% (n=5)(Tabela 4).

Ainda, observou-se a necessidade de verificar as diferenças dentro de cada categoria de conteúdo. Na categoria “conhecimentos gerais”, nos enunciados curtos com alternativas curtas, apenas 25% (n=2) das respostas foram acertadas. Nos enunciados curtos com alternativas longas e nos longos com alternativas curtas, as taxas de acerto permaneceram baixas, em 28,57% (n=4) e 25% (n=1), respectivamente. (n=2)(Tabela 4).

Na categoria "raciocínio diagnóstico", notou-se uma distribuição quase idêntica entre acertos e erros. Foram analisadas 22 questões de enunciados longos e alternativas curtas, resultando em uma taxa de acerto de 54,54% (n=12).

Já na categoria de “conduta”, nos enunciados longos com alternativas longas, 62,5% (n=5) das respostas estavam corretas. Mesmo nos enunciados longos com alternativas curtas, embora haja uma taxa de acerto menor, de 33,33% (n=1), a incidência de erros desse padrão, de 66,67% (n=2), é menor em relação às categorias de conteúdo anteriores (Tabela 4).

DISCUSSÃO

Enquanto os resultados se mostram positivos em certos aspectos, como quando utiliza-se enunciados longos com alternativas longas (54,55% de acertos) especialmente em questões sobre conduta (62,5% de acertos), se mostram muito negativos quando opta-se por enunciados curtos e alternativas longas (68,75% de erros). Além disso, também fica nítida a variação de desempenho em todo o estudo quando compara-se a taxa de acertos em enunciados longos (48,78%) com porcentagem de acertos em enunciados curtos (34,61%) - uma diferença de quase 15% quando o comando da questão é menor que 79 caracteres.

Esse decréscimo pode corresponder ao que se tem de conceito sobre uma das armadilhas comuns na escritas de um prompt (a armadilha da falta de contexto) (Giray, 2023)[7]. De acordo com Giray (2023) para corrigir isso, você deve aumentar o prompt incorporando dicas contextuais relevantes. Ao concretizar o prompt e especificar o contexto, você fornece ao modelo uma compreensão mais clara do escopo e do propósito da pergunta, permitindo-lhe gerar respostas mais precisas e abrangentes.

Esses resultados evidenciam a importância da engenharia de prompts, responsável pela fabricação do comando da questão, como um aspecto fundamental para a boa utilização de inteligências artificiais, no caso o GPT-3.5. Logo, a qualidade da resposta é diretamente proporcional a quantidade de informações com significância fornecidas e da qualidade de construção do prompt.

Na categoria "raciocínio diagnóstico", observou-se uma distribuição muito parecida entre acertos e erros das questões de enunciados curtos com alternativas longas e de enunciados longos com alternativas curtas. No primeiro caso, 50% das respostas foram corretas, enquanto no segundo caso, a precisão aumentou para 54,54%. Ainda assim, a predominância de respostas equivocadas prevaleceu nas categorias "conhecimentos gerais" e "raciocínio diagnóstico" .

Ademais, cabe destacar como limitações do estudo, o número de questões (n= 66 questões), a seleção através de um banco de dados restrito, a não escolha do gabarito de cada questão por parte dos autores do artigo e a subjetividade na classificação de grupos.

De maneira geral, a média aproximada obtida de 43% de acertos nos levam a questões como: Até que ponto o usuário pode "confiar" no GPT-4 ou o leitor precisa gastar tempo verificando a veracidade do que ele escreve? (Lee,2023)[8] E conclusões como: a adoção deste chatbot de IA deve ser conduzida com extrema cautela, considerando as suas potenciais limitações. (Sallam,2023)[9] - visto isso, a necessidade de cautela ao cogitar a utilização desse Chatbot para a prática médica é um consenso.

Portanto, antes de implementar o ChatGPT, as potenciais limitações e considerações éticas precisam ser cuidadosamente avaliadas e abordadas.

Este estudo sugere que a utilização do modelo de Inteligência Artificial GPT-3.5 na resolução de questões de múltipla escolha ortopédicas em português seja inconsciente. Contudo, vale ressaltar que perante a grande velocidade da evolução das Inteligências Artificiais novos estudos serão necessários para reavaliar a credencial das novas versões do ChatGPT e outras IAs.

CONCLUSÃO

A união dos modelos chatbot de inteligência artificial como o GPT-3.5 na prática médica ortopédica e acadêmica aparenta não estar pronta. Por mais que a média percentual de respostas corretas em determinados conjuntos ultrapasse a maioria, em geral não se observou tal resultado. Os ortopedistas, residentes e acadêmicos de medicina devem estar cientes da limitação que hoje possui o Chat GPT 3.5.

REFERÊNCIAS

1. SILVA, J. A. S.; MAIRINK, C. H. P. **Inteligência artificial: aliada ou inimiga. LIBERTAS: Rev. Ciênci.** Soc. Apl., Belo Horizonte, v. 9, n. 2, p. 64-85, ago./dez. 2019.

2. Lobo LC. **Inteligência artificial, o futuro da medicina e a educação médica**. Rev Bras Educ Med. 2018;42:3-8. doi: 10.1590/1981-52712015v42n3RB20180115EDITORIAL1.

3. Ramos ASM. **Inteligência artificial generativa baseada em grandes modelos de linguagem: ferramentas de uso na pesquisa acadêmica.** 2023.

4. Tang S, Yang X, Shajudeen P, Sears C, Taraballi F, Weiner B, Tasciotti E, Dollahon D, Park H, Righetti R**. A CNN-based method to reconstruct 3-D spine surfaces from US images in vivo**. Med Image Anal. 2021 Dec;74:102221. doi: 10.1016/j.media.2021.102221. Epub 2021 Sep 1. PMID: 34520960.

5. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, Stewart M, Collins GS, Furniss D. **Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis.** Radiology. 2022 Jul;304(1):50-62. doi: 10.1148/radiol.211785. Epub 2022 Mar 29. PMID: 35348381; PMCID: PMC9270679.

6. Rigamonti L, Albrecht UV, Lutter C, Tempel M, Wolfarth B, Back DA; **Working Group Digitalisation. Potentials of Digitalization in Sports Medicine: A Narrative Review. Curr Sports Med Rep.** 2020 Apr;19(4):157-163. doi: 10.1249/JSR.0000000000000704. PMID: 32282462.

7. Giray, L**. Prompt Engineering com ChatGPT: um guia para escritores acadêmicos.** Ann Biomed Eng 51 , 2629–2633 (2023). https://doi.org/10.1007/s10439-023-03272-4

8. Lee P, Bubeck S, Petro J. **Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med.** 2023 Mar 30;388(13):1233-1239. doi: 10.1056/NEJMsr2214184. PMID: 36988602.

9. Sallam M. **ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare** (Basel). 2023 Mar 19;11(6):887. doi: 10.3390/healthcare11060887. PMID: 36981544; PMCID: PMC10048148.