O GLOBO
Publicado em 10/02/2026 - 10:14

Pacientes raramente extraem informação médica útil da IA, mostram estudos

Mesmo sistemas capazes de responder questões complexas podem induzir ao erro quando consultados por leigos; linguagem usada por usuário pode comprometer resultado

Por Rafael Garcia — São Paulo

A qualidade da informação médica fornecida por ferramentas de inteligência artificial populares cai muito quando elas são usadas por pessoas comuns e não por cientistas, mostra um novo estudo.

No trabalho, publicado nesta segunda-feira (9), pesquisadores mostraram que o conteúdo da pergunta submetida aos chamados "grandes modelos de linguagem" (LLMs) afeta muito o resultado da resposta obtida.

Os autores do trabalho decidiram fazer o teste depois que esses serviços de IA, particularmente o ChatGPT e o Llama, começaram a ganhar certa reputação de "doutores virtuais" por acertarem questões usadas em provas de proficiência para médicos.

Na nova pesquisa, porém, cientistas da Universidade de Oxford mostraram que leigos não conseguem extrair esse desempenho dos chatbots por conta própria. E pior, com frequência a resposta obtida os induz ao erro na tomada de decisões.

Publicado na revista "Nature Medicine", o trabalho descreve um experimento online com participação de mais de 1.200 voluntários. A ideia era ver como eles se saíam, em média, ao tentar decifrar um problema de saúde apresentado.

Cada um dos participantes recebia uma descrição geral de sintomas correspondente a uma condição médica. Entre os cenários usados estavam casos de cólica renal, rinite alérgica, embolia pulmonar e outras condições bem conhecidas pela medicina.

A missão dos voluntários era conseguir obter um diagnóstico genérico do problema e descobrir como proceder (ir ao pronto-socorro, marcar uma consulta, tomar um

medicamento etc.). Parte deles podia usar chatbots de IA, e outra parte só tinha acesso a fontes de informação tradicionais na internet.

Para efeito de comparação, os cientistas também realizaram os testes consultando os LLMs diretamente, sem delegar a tarefa a voluntários como intermediários.

Apesar de os pesquisadores conseguiam extrair informações corretas da IA em 94,9% dos casos e de obterem a indicação do procedimento correto em 56,3% deles, o desempenho caia drasticamente nos outros cenários.

Se a tarefa era realizada via IA por pessoas leigas, taxa de acerto caía para 34,5%, e a tomada de atitude correta só ocorria 44,2% das vezes.

As pessoas que tinham acesso só a fontes de informação tradicionais, por fim, tinham quase duas vezes mais chances de acertar o diagnóstico por conta própria, quando comparadas aos voluntários que usaram as ferramentas de IA.

Apesar do sucesso dessa tecnologia em condições controladas, seu fraco desempenho no mundo real levantou preocupações. Já foi sugerida a possibilidade de a IA servir como uma espécie de ferramenta de cuidados primários, para triar pacientes, mas o novo estudo indica que é cedo demais para isso.

"Nossos resultados destacam os desafios para o emprego público de LLMs em cuidados diretos ao paciente", escreveram os autores do estudo, liderados pelo cientista Andrew Bean. "Apesar de os LLMs em si apresentarem alta proficiência na tarefa, a combinação de LLMs e usuários humanos não se mostrou superior ao grupo de controle em avaliar a gravidade clínica e apresentou desempenho inferior na identificação de condições relevantes."

No trabalho, os cientistas não investigam diretamente os motivos do fracasso, mas já têm algumas pistas. O caminho, afirmam, não é o de culpar os usuários por "não saberem usar" LLMs da melhor forma.

"Observamos casos tanto de participantes que forneceram informações incompletas quanto de profissionais de saúde que interpretaram erroneamente as perguntas dos usuários, levando a esse resultado", afirmam.

A falha, aparentemente, tem relação com o fato de que a pergunta que um leigo digita em uma janela de prompt do ChatGPT não se parece com a questão de uma prova escrita de proficiência médica. Além disso, pacientes nem sempre estão atentos para alguns sintomas importantes de observar em cada contexto.

Jargão enganador

Um segundo estudo, também publicado na segunda-feira (9), trouxe mais indícios de que parte do problema está na caixa de perguntas.

Nesse outro trabalho, liderado pela escola médica do hospital Mount Sinai, de Nova York, os cientistas testaram o quanto as ferramentas de IA conseguem identificar falácias médicas.

Eles fizeram o teste de duas maneiras. Na primeira delas forneciam aos LLMs textos equivocados tirados diretamente da plataforma de rede social Reddit. Eram frases comprovadamente incorretas dizendo, por exemplo, que Tylenol causa autismo ou que mamogramas causam câncer.

Numa segunda rodada, submeteram às ferramentas de IA tiradas diretamente de prontuários e outros documentos médicos. Nesse caso, os equívocos foram criados propositalmente pelos cientistas. Num dos prontuários, por exemplo, afirmaram que tomar leite fria era uma maneira eficaz de estancar um sangramento interno.

Mais de 3,4 milhões de pronts foram usados no teste. O resultado é que a IA textual falhou em identificar informações corretas em um grande número de casos.

Foram testados 20 LLMs diferentes, e em média eles foram incapazes de detectar 31,7% das informações incorretas. Os modelos "grandes", como o ChatGPT 4.0, se saíram melhor, deixando passar 10,6% dos erros, ainda uma proporção preocupante. O pior dos modelos foi o Gemma-3-4B-it, que só identificou um terço dos erros.

Os cientistas notaram um outro padrão interessante nos dados: quando os chatbots eram desafiados com conteúdo do Reddit, eles identificavam erros mais facilmente do que quando eram submetidos aos prontuários médicos adulterados, escritos em jargão médico.

"Esses resultados mostram que os LLMs atuais ainda absorvem falsidades médicas nocivas, especialmente quando redigidas em linguagem clínica autoritária" escreveram os autores do estudo, liderados pelo cientista Mahmud Omar. "Paradoxalmente, os LLMs se otrnam menos vulneráveis ??quando as mesmas afirmações são apresentadas em estilos que induzem à falácia lógica."

O fraseado de respostas no Reddit frequentemente vem acompanhado de chavões como "todos sabem que..." ou "um famoso médico disse que...". Esse estilo retórico, diz, parece ter servido à IA como um sinal de alerta para informação

errônea.

O estudo de Omar e seus colegas do Mount Sinal foi publicado pela revista médica "The Lancet Digital Health". O resultado, dizem os pesquisadores, é mais um sinal de que ainda é cedo para depositar confiança nos chatbots de IA como fonte de informação médica.

https://oglobo.globo.com/saude/ciencia/noticia/2026/02/10/pacientes-raramente-extraem-informacao-medica-util-da-ia-mostram-estudos.ghtml

Veículo: Online -> Portal -> Portal O Globo - Rio de Janeiro/RJ