ESTADÃO
Publicado em 20/02/2026 - 10:12

IA dá conselhos de saúde, mas estudo revela: muitos estão errados

Em parte, o problema tem a ver com a forma como os usuários fazem as perguntas

Por Teddy Rosenbluth (The New York Times)

Um novo estudo publicado recentemente trouxe uma visão preocupante sobre se os chatbots de IA, que rapidamente se tornaram uma importante fonte de informações de saúde, são de fato bons em fornecer aconselhamento médico ao público em geral.

O experimento constatou que os chatbots não eram melhores do que o Google — já considerado uma fonte falha de informações médicas — para orientar usuários rumo aos diagnósticos corretos ou ajudá-los a decidir o que deveriam fazer em seguida. E a tecnologia apresentou riscos específicos, às vezes fornecendo informações falsas ou mudando drasticamente suas recomendações dependendo de pequenas variações na forma como as perguntas eram formuladas.

Nenhum dos modelos avaliados no experimento estava “pronto para uso em atendimento direto a pacientes”, concluíram os pesquisadores no artigo, que é o primeiro estudo randomizado desse tipo.

Nos três anos desde que os chatbots de IA foram disponibilizados ao público, perguntas sobre saúde se tornaram um dos temas mais comuns feitos pelos usuários.

Alguns médicos veem regularmente pacientes que consultaram um modelo de IA para obter uma primeira opinião. Pesquisas mostraram que cerca de um em cada seis adultos usou chatbots para encontrar informações de saúde pelo menos uma vez por mês. Grandes empresas de IA, incluindo Amazon e OpenAI, lançaram produtos especificamente voltados para responder às dúvidas de saúde dos usuários.

Essas ferramentas despertaram entusiasmo por boas razões: os modelos já passaram em exames de licença médica e superaram médicos em problemas diagnósticos desafiadores.

Mas Adam Mahdi, professor do Oxford Internet Institute e autor sênior do novo estudo, suspeitava que essas perguntas médicas limpas e diretas não eram um bom indicativo de desempenho com pacientes reais. “A medicina não é assim”, diz. “A medicina é confusa, incompleta, estocástica.”

Então ele e seus colegas montaram um experimento. Mais de 1.200 participantes britânicos, a maioria sem formação médica, receberam um cenário clínico detalhado, com sintomas, informações gerais de estilo de vida e histórico médico. Os pesquisadores orientaram os participantes a conversar com o bot para descobrir os próximos passos adequados, como ligar para uma ambulância ou se tratar em casa. Foram testados chatbots comerciais como o ChatGPT, da OpenAI, e o Llama, da Meta.

Os pesquisadores descobriram que os participantes escolheram o curso de ação “correto” — previamente definido por um painel de médicos — menos da metade das vezes. E os usuários identificaram corretamente as condições, como cálculos biliares ou hemorragia subaracnoide, cerca de 34% das vezes.

Eles não tiveram desempenho melhor do que o grupo de controle, que recebeu a orientação de realizar a mesma tarefa usando qualquer método de pesquisa que normalmente utilizariam em casa, principalmente buscas no Google.

O experimento não é uma janela perfeita para entender como os chatbots respondem perguntas médicas no mundo real: os usuários do estudo perguntaram sobre cenários fictícios, o que pode ser diferente de como interagiriam com os chatbots sobre a própria saúde, pondera Ethan Goh, que lidera a rede de avaliação de pesquisa em IA da Stanford University.

E como as empresas de IA frequentemente lançam novas versões dos modelos, os chatbots usados pelos participantes um ano antes durante o experimento provavelmente são diferentes daqueles com que os usuários interagem hoje. Um porta-voz da OpenAI disse que os modelos que atualmente alimentam o ChatGPT são significativamente melhores para responder perguntas de saúde do que o modelo testado no estudo, que desde então foi descontinuado. Eles citaram dados internos mostrando que muitos modelos novos têm muito menos probabilidade de cometer erros comuns, incluindo alucinações e falhas em situações potencialmente urgentes. A Meta não respondeu a um pedido de comentário.

Ainda assim, o estudo lança luz sobre como interações com chatbots podem dar errado.

Quando os pesquisadores analisaram mais profundamente essas interações, descobriram que cerca de metade das vezes os erros pareciam resultar de falhas dos próprios usuários. Os participantes não inseriam informações suficientes ou os sintomas mais relevantes, e os chatbots acabavam dando recomendações com uma visão incompleta do problema.

Um modelo sugeriu a um usuário que as “dores fortes no estômago” que duraram uma hora poderiam ter sido causadas por indigestão. Mas o participante não havia incluído detalhes sobre a intensidade, localização e frequência da dor — todos fatores que provavelmente teriam levado o bot ao diagnóstico correto, cálculos biliares.

Em contraste, quando os pesquisadores inseriram o cenário médico completo diretamente nos chatbots, eles diagnosticaram corretamente o problema 94% das vezes.

Uma parte importante do que os médicos aprendem na faculdade é reconhecer quais detalhes são relevantes e quais devem ser descartados.

“Existe muita mágica cognitiva e experiência envolvidas em descobrir quais elementos do caso são importantes para inserir no bot”, explica Robert Wachter, chefe do departamento de medicina da University of California, San Francisco, que estuda IA na área da saúde.

Mas Andrew Bean, estudante de pós-graduação em Oxford e autor principal do artigo, afirma que o peso não deveria necessariamente recair sobre os usuários para formular a pergunta perfeita. Segundo ele, os chatbots deveriam fazer perguntas de acompanhamento, de maneira semelhante ao modo como médicos coletam informações dos pacientes.

“É realmente responsabilidade do usuário saber quais sintomas destacar, ou é em parte responsabilidade do modelo saber o que perguntar?”, questiona.

Essa é uma área que as empresas de tecnologia estão tentando melhorar. Por exemplo, os modelos atuais do ChatGPT têm cerca de seis vezes mais probabilidade de fazer uma pergunta de acompanhamento do que a versão anterior, de acordo com dados fornecidos por um porta-voz da OpenAI.

Mesmo quando os pesquisadores digitavam diretamente o cenário médico, descobriram que os chatbots tinham dificuldade em distinguir corretamente quando um conjunto de sintomas exigia atenção médica imediata ou cuidados não urgentes. Danielle Bitterman, que estuda interações entre pacientes e IA na Mass

General Brigham, informa que isso provavelmente ocorre porque os modelos são treinados principalmente com grandes volumes de livros médicos e relatos de casos, mas têm muito menos experiência com a tomada de decisão livre que médicos adquirem com a prática.

Em várias ocasiões, os chatbots também forneceram informações inventadas. Em um caso, um modelo orientou um participante a ligar para uma linha de emergência que não tinha dígitos suficientes para ser um número de telefone real.

Os pesquisadores também identificaram outro problema: até pequenas variações na forma como os participantes descreviam seus sintomas ou faziam perguntas mudavam significativamente as recomendações do bot.

Por exemplo, dois participantes do estudo tinham as mesmas informações iniciais — dor de cabeça intensa, sensibilidade à luz e rigidez no pescoço — mas descreveram o problema aos chatbots de maneira um pouco diferente.

Em um caso, o chatbot tratou a situação como algo leve que não exigia atenção médica imediata. Na outra resposta, o chatbot considerou os sintomas um sinal de problema grave e orientou o usuário a ir ao pronto-socorro. “Palavras muito, muito pequenas fazem diferenças muito grandes”, resume Bean.

Este texto foi traduzido com o auxílio de ferramentas de Inteligência Artificial e revisado por nossa equipe editorial. Saiba mais em nossa Política de IA.

https://www.estadao.com.br/saude/ia-da-conselhos-de-saude-mas-estudo-revela-muitos-estao-errados/

Veículo: Online -> Portal -> Portal Estadão