

Volume 3 Issue 1 Spring 2026

Hypothesis



PERSPECTIVES INSIGHTS & THOUGHT LEADERSHIP IN THE LIFE SCIENCES

Reliability of LLMs as Medical Assistants for the General Public: A Randomized Preregistered Study



By Andrew M. Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S. Ekanayaka, Lionel Tarassenko, Luc Rocher, and Adam Mahdi

INSIDE THIS ISSUE

HEALTH MUSINGS

The Hardest Sell

INSIGHTS AND ANALYSIS

ChatGPT Health Imagines an Ideal Patient—But That's Not Who Will Use It

PHYSICIAN'S PERSPECTIVE

I'm a Doctor Who Knows Better—But I Switched to an Inferior Drug to Save Money



Medical minds gather here.

As the largest independent medical publisher in Canada, our peer-reviewed open access scientific journals are a practical resource for Canadian healthcare practitioners. We currently publish specialty journals in the areas of allergy & immunology, dermatology, hematology, ophthalmology, diabetes & endocrinology, gastroenterology, primary care, women's health, rheumatology, oncology, respiratory and our press is constantly growing with new titles planned.



CONTENTS | Hypothesis

PERSPECTIVES INSIGHTS & THOUGHT LEADERSHIP IN THE LIFE SCIENCES

Volume 3 Issue 1 Spring 2026



07



19



26



30

- 04** Letter From Our Editor-In-Chief
- 05** Health Musings
The Hardest Sell
- 07** Spotlight Article
Reliability of LLMs as Medical Assistants for the General Public:
A Randomized Preregistered Study
- 19** Physician's Perspective
I'm a Doctor Who Knows Better—But I Switched to an
Inferior Drug to Save Money
- 22** Calendar of Events
- 24** Product Newswire
Information on the Latest Drug Approvals and
Reimbursement Milestones
- 26** Insights and Analysis
ChatGPT Health Imagines an Ideal Patient—But That's
Not Who Will Use It
- 30** Research and Findings
In Case You Missed It...
- 34** People on The Move
Who's Doing What and Who's Going Where

Editorial Team

Managing Editor, President & CEO
Rohit Khanna

Editor-in-Chief
Rohit Khanna

Senior Writer
Wendy Glauser

Senior VP, Strategy
John Andonoff

Director, Client Services
Lou-Anne Gaudino

Senior Project Manager
Treena Cadiz

Vice President, Consulting Services
Alanna Fox

Senior Art Director
Kevin Buist

Art Director
Emily Redmond

Director of Digital Design
Krishna Shenoy

Advertising Opportunities
john@catalytichealth.com

Questions & Submissions
info@catalytichealth.com

LinkedIn
<https://www.linkedin.com/showcase/hypothesis-magazine/>

Letter From Our Editor-in-Chief

Welcome back for our third year of publication! We're really thankful to be able to continue this journey with you and to curate some great healthcare-related stories that we think matter. If you like what we're doing, please share the journal link with a friend or colleague: www.hypothesismag.com.

It takes a (small) village of dedicated team members at Catalytic Health to produce this magazine three times per year. But it also takes advertising support, and we need to thank our loyal advertisers: The Bedford Group, TRANSEARCH, Impres Pharma, and IQVIA.

Look, you're going to see a lot about AI in this (and probably every) issue going forward for the foreseeable future. We think it's important. But most importantly, you've told us that you want that content. I had the chance to speak with a couple of dozen readers and the feedback have been the same: we love the "made in Canada" feel of the magazine, but we need to stay on top of trends too.

We hear you.

There's no bigger trend than AI and its impact across healthcare promises to be head spinning over the years to come and we've got you covered. Some of the stuff is dense (see [page 7](#)) and some of it is a little easier to digest (see [page 26](#)). But, in the end, it all helps us understand the potential promise and pitfalls of AI in our industry.

Take a look at the "In Case You Missed It" section in this issue. It's an eye-opener. Or is it? We all know how polarized society has become over the last 10 years, but did we think that it was this bad? And did we think it affected our views on science? After COVID-19, nothing about the polarization of scientific views (North or South of the border) surprises me anymore.

And don't miss the "Physician's Perspective" section either. We may not have the exact same model here in Canada (high-deductible health plans), but one of the take home messages from this may be that physician and patient discrete choices aren't that different when the physician is the patient.

We hope you enjoy this issue and that you'll share it with your peers. We welcome your feedback!



Editor-in-Chief
Hypothesis Magazine



Rohit Khanna, MBA, MSc, MPH

Rohit Khanna is the Managing Director of Catalytic Health, a leading life-sciences communication, publishing, and strategy firm. He holds a B.A. from McGill University, an M.B.A. from Queen's School of Business, an MSc. from the London School of Economics & Political Science and a Master of Public Health in Epidemiology from Harvard School of Public Health. His first book entitled *Misunderstanding Health: Making Sense of America's Broken Health Care System* was published in October 2021 by Johns Hopkins University Press. He can be reached at rohit@catalytichealth.com

The Hardest Sell

Consumer choice seems obvious. Except when it comes to healthcare advertising.

By Rohit Khanna, MBA, MSc, MPH



I'm generally in favour of anything that gives consumers more choice and more information when deciding what to do or what to buy.

But when it comes to healthcare, it gets complicated as it always seems to do.

In 2014, Fain and Alexander **laid out** the pros and cons of direct-to-consumer (DTC) advertising in their paper, which included obvious benefits of increased awareness of treatment choice and enhanced physician-patient conversations about disease management. The downsides are equally obvious: overprescribing medicines and inappropriate use.

More recently, Alpert and Sood **published** a paper in *JAMA Health Forum* last year, which suggests that the case for banning DTC advertising is not as strong as it may initially appear. Alpert and Sood present the prevailing view as follows: "critics argue that DTC advertisements are misleading and promote low-value and high-cost drugs, leading to inappropriate use and encouraging patients to choose advertised medications vs cheaper, similarly effective generics. There are also concerns that DTC advertising undermines trust in the physician-patient relationship by creating disagreements over treatment choices." But they debunk these concerns with insights and data from research on the topic, which as you may have guessed



is not randomized or robust in its methodology, but nonetheless presents a compelling point of view.

We recently completed a DTC campaign for a client (and are about to embark on another). The process was not unexpected. We briefed with the client about the objectives, the target audience, and the potential platforms for the ensuing media buy. We developed creative content and a media plan with costs and timing.

Then we submitted the creative content to Advertising Standards Canada (ASC). Here's the response we got: *In Canada, health product advertising is regulated under the Food and Drugs Act and Food and Drug Regulations. Relevant sections for prescription drug advertising include the following: Section C.01.044 of the Food and Drug Regulations prohibits advertising of prescription drugs to consumers, beyond mention of name, price, and quantity of the drug.*

A few of our creative submissions might have been 'on the edge'. But many were not. And they all got rejected with the same 'cut and paste' response. It seems that a simple statement or visual, with a call to action (ask your healthcare professional), and a URL (please visit www.XXXXXX.xa for more information) requires a Watergate-like level of scrutiny and analysis. We need to have public hearings. We need to reconvene Parliament for a vote.

Of course, I am exaggerating.

In an age where Dr Google and internet health information-related searches continue to thrive. In an age where conventional and speciality TV channels have been replaced by influencers on social media. In an age where AI platforms are handling tens of millions of healthcare-related prompts and queries every day. And in an age where close to 90% of Canadians have a smartphone and can literally find out the product's indication in seconds.

We are still stuck on name, price, and quantity.

Do we really still need to be careful about making sure we say "Ozempic. Ask your doctor." We can't say "Talk to your doctor about whether Ozempic is right for your Type 2 diabetes". Really? We're so convinced that physicians in this country are going to write prescriptions for Ozempic without doing a proper workup and that they're going to give patients glucagon-like peptide-1 receptor (GLP-1 RAs) even though metformin, insulin, or an sodium-glucose co-transporter-2 inhibitors (SGLT-2i) might be a better choice? We really think that shared decision-making is a euphemism for "I'm just going to give the patient what they ask for." Or do we really believe that public and private health plans won't act as natural 'gatekeepers' for any potential abuse associated with DTC advertising. As if asking for a GLP-1RA for your Type 2 diabetes is going to make Canada Life or Sun Life or Great West Life ignore any 'step through' therapies before getting to a GLP-1RA.

It's time to give up this silly, outdated policy of restricting DTC advertising to name, price, and quantity. We can do better. We can sit down and figure out a better way to ensure that consumers are protected from (potential) confusing and misleading advertising without pretending that there isn't a single person in this country who cannot find out what a drug is indicated for within seconds of seeing an ad for it.

We've gotten into this echo chamber mentality of believing that this is the only way to 'protect' consumers who are incapable of distinguishing between high-value and low-value information. And that doctors and insurance companies won't act as natural gatekeepers.

It's a hard sell. Or is it? ✨



Reliability of LLMs as Medical Assistants for the General Public: A Randomized Preregistered Study

By Andrew M. Bean¹, Rebecca Elizabeth Payne^{2,3,4},
Guy Parsons^{1,5}, Hannah Rose Kirk¹, Juan Ciro⁶, Rafael Mosquera-Gómez^{7,8},
Sara Hincapié M^{7,8}, Aruna S. Ekanayaka⁹, Lionel Tarassenko¹⁰,
Luc Rocher^{1,11} and Adam Mahdi^{1,11}



This article originally appeared in Nature Medicine at <https://www.nature.com/articles/s41591-025-04074-y> on February 9th, 2026.

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Affiliations: ¹Oxford Internet Institute, University of Oxford, Oxford, UK. ²Nuffield Department of Primary Health Care Sciences, University of Oxford, Oxford, UK. ³North Wales Medical School, Bangor University, Bangor, UK. ⁴Betsi Cadwaladr University Health Board, Ysbyty Gwynedd, Bangor, UK. ⁵National Health Service, London, UK. ⁶Contextual AI, Mountain View, CA, USA. ⁷MLCommons, San Francisco, CA, USA. ⁸Factored AI, Palo Alto, CA, USA. ⁹Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. ¹⁰Institute of Biomedical Engineering, University of Oxford, Oxford, UK. ¹¹ These authors jointly supervised this work: L. Rocher, A. Mahdi. e-mail: adam.mahdi@oii.ox.ac.uk

Global healthcare providers are exploring the use of large language models (LLMs) to provide medical advice to the public. LLMs now achieve nearly perfect scores on medical licensing exams, but this does not necessarily translate to accurate performance in real-world settings. We tested whether LLMs can assist members of the public in identifying underlying conditions and choosing a course of action (disposition) in ten medical scenarios in a controlled study with 1,298 participants. Participants were randomly assigned to receive assistance from an LLM (GPT-4o, Llama 3, Command R+) or a source of their choice (control). Tested alone, LLMs complete the scenarios accurately, correctly identifying conditions in 94.9% of cases and disposition in 56.3% on average. However, participants using the same LLMs identified relevant conditions in fewer than 34.5% of cases and disposition in fewer than 44.2%, both no better than the control group. We identify user interactions as a challenge to the deployment of LLMs for medical advice. Standard benchmarks for medical knowledge and simulated patient interactions do not predict the failures we find with human participants. Moving forward, we recommend systematic human user testing to evaluate interactive capabilities before public deployments in healthcare.

Recent breakthroughs in artificial intelligence (AI) research have the potential to democratize healthcare by expanding access to medical knowledge, bringing care closer to patients. The development of large language models (LLMs) such as OpenAI's ChatGPT could enable individuals to perform preliminary health assessments, receive personalized medical guidance and manage chronic conditions without immediate clinician intervention. Testimonies of patients having

used LLMs to successfully diagnose their own conditions are now common¹.

Surveys indicate that a growing number of people are already turning to AI-powered chatbots for sensitive health-related inquiries, with one in six American adults consulting AI chatbots for health information at least once a month^{2,3}.

Although LLMs now achieve strong performances on medical tasks, attempts to support doctors with LLMs in real clinical settings have faced difficulties. On the one hand, LLM scores on medical knowledge benchmarks are now commensurate with passing the US Medical Licensing Exam⁴. LLM-generated clinical documents are rated as equivalent to or better than those written by doctors^{5,6}. On the other hand, excelling at medical tasks in silico does not translate to accurate performance in clinical settings under physician guidance. For instance, one study showed that radiologists assisted by AI did not perform better at reading chest X-rays than without AI assistance, and both performed worse than AI alone⁷. Another study showed that physicians assisted by LLMs only marginally outperformed unassisted physicians in diagnosis problems, and both performed worse than LLMs alone⁸. Providing doctors with highly capable AI systems is not enough to meaningfully assist them on important tasks⁹. Healthcare professionals often struggle to appropriately assess and incorporate AI-generated recommendations, limiting the benefits of AI assistance¹⁰⁻¹².

More promising and easier to deploy, LLM-powered chatbots have instead been suggested as a ‘new front door’ to healthcare for patients who lack medical expertise^{13,14}. As a first point of contact for healthcare support, they could be used to broaden access to medical expertise and support overburdened health systems¹⁴⁻¹⁶. Medical experts have had mixed opinions on the prospects of having LLMs directly advise patients, citing problems of oversight and liability¹⁷ but also the possible benefits of providing support outside of clinical settings^{18,19}. In response to this opportunity, private companies have made considerable efforts to create language models suitable for healthcare applications²⁰⁻²².

To understand whether LLMs can reliably support the general public and bring care closer to patients, we conducted a study with 1,298 UK participants. Each participant was tasked with identifying potential health conditions and a recommended disposition (course of action) in response to one of ten different medical scenarios. The scenarios were developed by a group of three doctors who unanimously agreed on the correct dispositions for each. The scenarios were then given to a distinct group of four doctors to provide differential diagnoses (see **Figure 1** for the study design).

We then randomly assigned participants to four experimental arms, with stratification based on

demographics to ensure that each group had a composition similar to the national adult population. Participants in three treatment groups were provided with an LLM (GPT-4o, Llama 3, Command R+) for assistance in identifying conditions and dispositions—providing us with a diverse set of models that could all be used for accessing medical information. Participants in the control group were instructed to instead use any methods they would typically employ at home.

First, despite selecting three LLMs that were successful at identifying dispositions and conditions alone, we found that participants struggled to use them effectively. To explain these findings, we examined the transcripts of participant interactions with LLMs. Across all transcripts, we found that LLMs usually suggested at least one relevant condition—but less often than when LLMs alone were provided the entire scenario and tasked to output the relevant condition. We observed cases both of participants providing incomplete information and of LLMs misinterpreting user queries leading to this outcome. Furthermore, participants did not consistently follow these recommendations, suggesting that the performance issues when pairing participants with LLMs may be attributed to human–LLM interaction failures.

Second, we show that evaluations on standard benchmarks—often used to ensure safety and reliability before deployment—are not able to predict human–LLM interaction failures. Medical knowledge is typically benchmarked using questions from medical licensing examinations⁴. We compiled benchmark questions from topics matched to our scenarios and compared LLM performance on these questions to performance in the corresponding interactive testing for each model and scenario. Performance in structured question-answering tasks, higher than in interactive testing as expected in 26 out of 30 instances, was largely uncorrelated to the interactive testing.

Finally, we show that simulations of user interactions with LLMs—a promising method for creating realistic benchmarks—also do not predict human–LLM interaction failures. Adapting the techniques used to simulate patient interactions with LLMs^{23,24}, we replicated our interactive testing by replacing each of our human participants with an LLM-simulated user. Compared to our interactive testing with human participants, LLMs scored better on simulations. Crucially, the distribution of results did not reflect

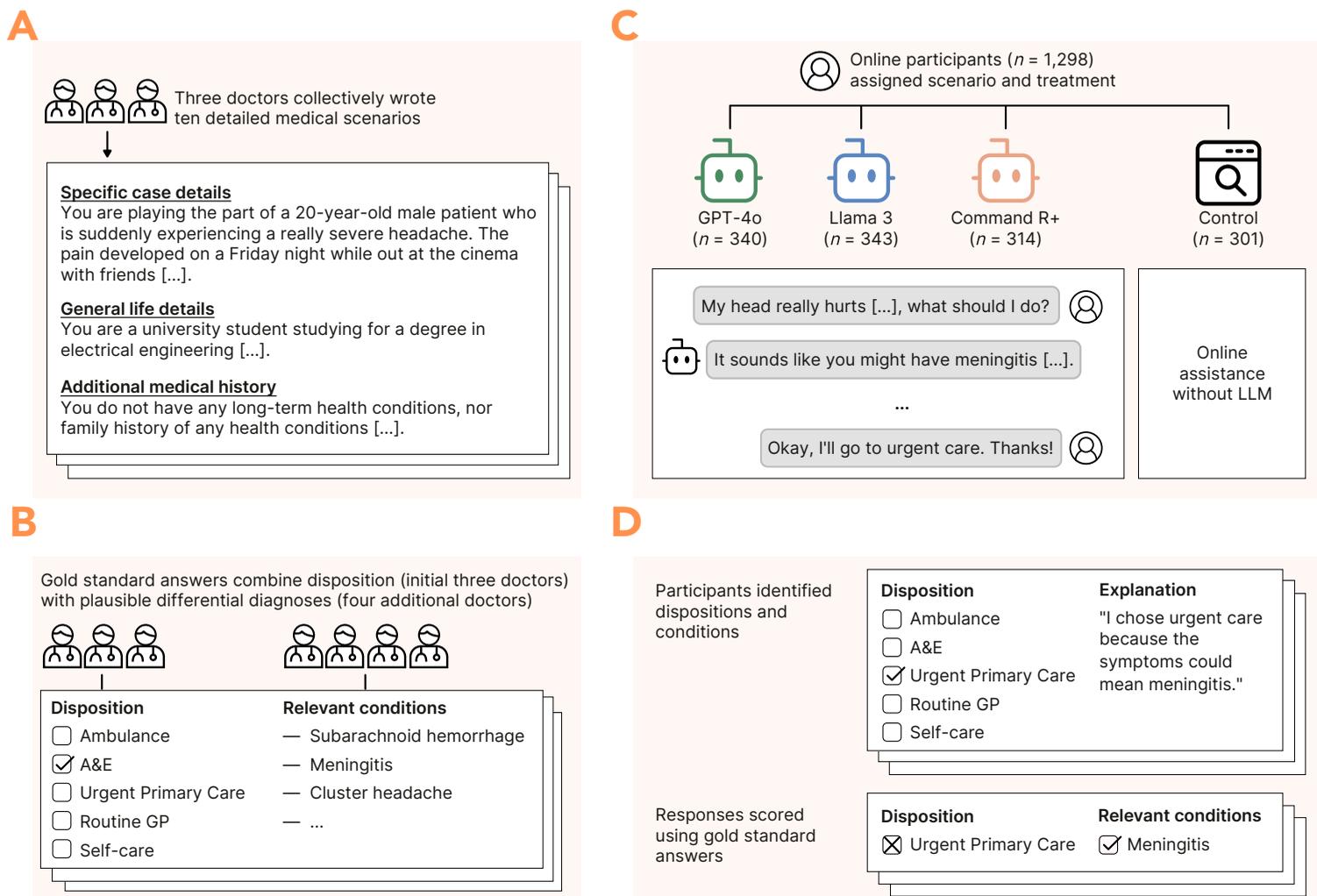


Figure 1. Study design. **a)** Three doctors drafted ten medical scenarios, iteratively revising them until they reached unanimous agreement about the best disposition on a five-point scale from self-care to ambulance. **b)** Four additional doctors read the scenarios and provided differential diagnoses, which were combined to form gold-standard lists of relevant conditions. **c)** We recruited 1,298 participants and randomly assigned them to one of four experimental conditions. Each participant was randomly allocated one of ten medical scenarios. The treatment groups conversed with an LLM to help assess the scenarios. The control group was permitted to use any method, with most participants using internet search or their own knowledge. **d) Top:** participants then chose a disposition and identified medical conditions that motivated their choice. Participants completed two scenarios, until a total of 600 examples were collected for each experimental condition. **Bottom:** we evaluated each participant's responses using the gold-standard answers.

human variability, and results were only weakly correlated to the interactive testing.

Taken together, our findings suggest that the safe deployment of LLMs as public medical assistants will require capabilities beyond expert-level medical knowledge. Despite strong performance on medical benchmarks, providing people with current generations of LLMs does not appear to improve their understanding of medical information. Fixing that will require identifying why humans fail when interacting with LLM-based tools—for example, aversion to technology and implicit biases against algorithms^{10,11,25} or LLM affordances undermining trust in interactions²⁶—and, crucially, how to design more reliable and deterministic conversational

LLM-based tools in high-risk settings. Importantly, our work shows that solving these challenges will require moving away from benchmarks and simulations to systematically conduct safety testing with diverse, real users; this is crucial to ensure the reliability of AI systems and enable potential benefits for healthcare.

Results

To assess the risks of the public using LLMs for medical advice, we conducted a randomized study where we asked participants to make decisions about a medical scenario as though they had encountered it at home (**Figure 1**). We created ten scenarios where a patient must decide whether and how to access professional

medical treatment. In each scenario, participants chose the best disposition on a five-point scale, ranging from staying home to calling an ambulance, and listed the medical conditions they had considered that led to their choice. We scored the selected disposition based on whether it matched the answer given by the three physicians involved in drafting the case. We scored the listed medical conditions based on whether they appeared in a gold-standard list of relevant conditions generated by four physicians unfamiliar with the scenarios.

We recruited 1,298 participants living in the UK and over the age of 18 (**Figure 1b**). Participants were randomly assigned to one of three treatment groups or the control to provide a maximum of two responses, with the sample population for each experimental condition stratified to reflect the demographics of the UK. Data collection continued until 600 responses were collected for each experimental condition. Participants in the treatment groups interacted with one of GPT-4o, Llama 3 or Command R+ at least once per scenario and as many times as they desired to help them decide how to respond to the questions. We chose these models to represent widely used LLMs as well as the approach of using internet search to augment responses. For the control group, we instructed participants to use any assistance they would typically use at home (for example, internet search).

Task validation

As a validation of the potential usefulness of the LLMs for addressing our specific set of ten scenarios, we provided the scenarios and questions directly to the models and sampled 60 responses per model per scenario. The models were able to suggest at least one relevant condition in 94.7% of cases for GPT-4o, 99.2% of cases for Llama 3 and 90.8% of cases for Command R+. The models' accuracy in recommending dispositions was 64.7% for GPT-4o, 48.8% for Llama 3 and 55.5% for Command R+ (**Figure 2a**). Overall, these scores indicate that the models have the ability to provide useful medical information on these tasks, performing better than random guessing between the five choices of disposition or choosing medical conditions from our gold-standard lists to recommend at random and consistent with the strong performance of these models on other medical benchmarks^{6,22,27,28}.

Experimental Performance

Figure 2b shows that participants using LLMs were significantly less likely than those in the control group to correctly identify at least one medical condition relevant to their scenario ($\chi^2(1)$, $n_1 = n_2 = 600$, $P < 0.001$ for all three models) and identified fewer relevant conditions on average (GPT-4o, 0.42–0.54; Llama 3, 0.39–0.50; Command R+, 0.34–0.43; control, 0.55–0.67; bootstrap 95% confidence interval (CI) with 1,000 resamples). Participants in the control group had 1.76 (95% CI = 1.45–2.13) times higher odds of identifying a relevant condition than the aggregate of the participants using LLMs. They were also 1.57 (95% CI = 1.28–1.92) times more likely to identify conditions from the more serious 'red flag' list. Participants using LLMs did not have statistically significant differences in disposition accuracy from the control group (GPT-4o, $\chi^2(1) = 0.17$, $P = 0.683$; Llama 3, $\chi^2(1) = 0.34$, $P = 0.560$; Command R+, $\chi^2(1) = 0.03$, $P = 0.861$; $n_1 = n_2 = 600$; **Figure 2b**). The overall correct response rate of $43.0\% \pm 2.0\%$ exceeds a random guessing baseline of 20%, but most participants still chose an incorrect disposition. Participants using LLMs tended to underestimate the acuity of their conditions, as did the control group (Mann–Whitney U $n_1 = n_2 = 600$, $P < 0.001$ for all experimental conditions). Users of GPT-4o and Llama 3 had an observed tendency toward higher estimates of clinical acuity than the control group, but this result was not significant (Mann-Whitney U $n_1 = n_2 = 600$; GPT-4o, $F = 0.536$, $P = 0.023$; Llama 3, $F = 0.529$, $P = 0.072$; Command R+, $F = 0.514$, $P = 0.366$, unadjusted).

Participants using LLMs consistently performed worse than when the LLMs were directly provided with the scenario and task (**Figure 2**). For identifying relevant conditions, all three treatment groups performed worse than their corresponding models without human interaction ($\chi^2(1)$, $P < 0.001$, $n_1 = n_2 = 600$). In the case of disposition, GPT-4o and Command R+ performed better than any group of participants using LLMs ($\chi^2(1)$, $P < 0.001$, $n_1 = n_2 = 600$), and the observed mean was higher for Llama 3 as well but not statistically significant ($\chi^2(1) = 2.44$, $P = 0.118$, $n_1 = n_2 = 600$). Strong performance from the LLMs operating alone is not sufficient for strong performance with users.

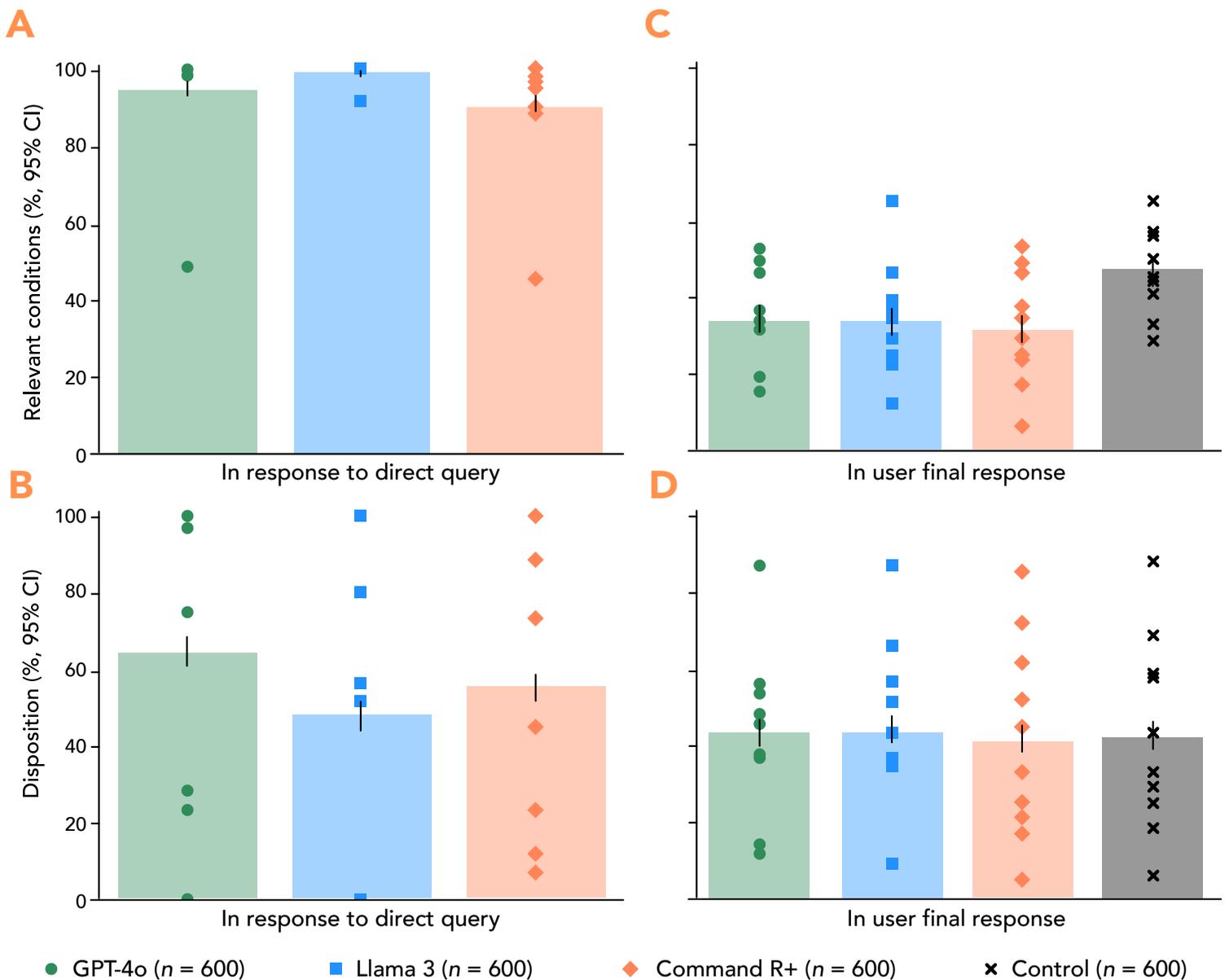


Figure 2. Performance of LLMs alone and with users. **a)** The performance on the LLMs when directly prompted to complete each task alone. Top, the proportion of LLM responses that identified relevant conditions. Bottom, the proportion of LLM responses correctly identifying the best disposition. **b)** The performance of participants across the four experimental conditions. **Top:** the proportion of participant responses that identified relevant conditions. **Bottom:** the proportion of participant responses correctly identifying the best disposition. The control group was significantly better than those using LLMs at identifying relevant conditions. Differences in disposition accuracy were not statistically significant. Data are presented as mean values with unadjusted 95% CIs for proportions. Markers indicate means for each scenario. Using LLMs worsened or did not improve participant performance on these tasks relative to using traditional resources, and the models consistently performed better without user interaction.

Performance in User Interactions

To isolate the role of user interactions, we compared the performance in identifying relevant conditions with different degrees of user involvement. In the user interactions, GPT-4o mentioned a relevant condition in $65.7 \pm 6.2\%$ of cases, Llama 3 in $67.0 \pm 6.1\%$ and Command R+ in $73.2 \pm 5.7\%$ (**Figure 3**). Each of these was significantly lower than the performance of the LLMs alone (**Figure 2**) and suggests that necessary

information about the scenario was not communicated between the user and the model. Despite these correct suggestions appearing in the conversations, users did not consistently include them in the final responses, indicating a second breakdown in communication between the model and user (**Figure 3**).

To measure whether participants received accurate suggestions from LLMs, we searched each conversation to identify which medical conditions were mentioned

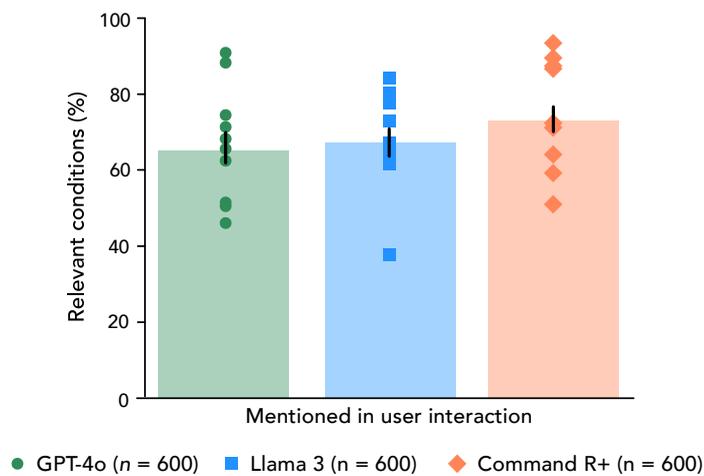


Figure 3. Identification of relevant conditions in interaction. The proportion of responses where at least one condition from the gold-standard list (Supplementary Table 11) was mentioned during the interaction between the participants and LLMs. Data are presented as mean values with unadjusted 95% CIs for proportions. Markers indicate means for each scenario.

(‘Condition extraction’ in Methods). We found that, on average, LLMs suggested 2.21 (2.12–2.32 95% CI) possible conditions per interaction, of which only 34.0% (32.3–35.9% 95% CI) were correct. After their interactions with LLMs, participants were asked to list all relevant conditions and listed 1.33 (1.28–1.38 95% CI) on average. We found that user final responses had only slightly better precision, 38.7% (36.3–41.4% 95% CI), than the combination of all the intermediate conditions mentioned by LLMs. This indicates that participants may not be able to identify the best conditions suggested by LLMs.

To better understand the mechanisms that lead to lower performance in human–LLM interactions, we analyzed a random selection of 30 interactions, one for each combination of model and scenario. For each selected interaction, we read the interaction transcript and recorded **1**) whether the user had provided sufficient information to correctly identify the condition in their initial message, **2**) whether the user had provided sufficient information over the course of the interaction, **3**) the accuracy of any suggestions made by the model and **4**) whether the user ultimately followed the recommendation of the model. We also noted interactions that demonstrated unique dynamics between LLM and user.

Overall, users often failed to provide the models with sufficient information to reach a correct recommendation. In 16 of 30 sampled interactions, initial messages contained only partial information

(see **Extended Data Table 1** for a transcript example). In 7 of these 16 interactions, users mentioned additional symptoms later, either in response to a question from the model or independently.

LLMs generated several types of misleading and incorrect information. In two cases, LLMs provided initially correct responses but added new and incorrect responses after the users added additional details. In two other cases, LLMs did not provide a broad response but narrowly expanded on a single term within the user’s message (‘pre-eclampsia’ and ‘Saudi Arabia’) that was not central to the scenario. LLMs also made errors in contextual understanding by, for example, recommending calling a partial US phone number and, in the same interaction, recommending calling ‘Triple Zero’, the Australian emergency number. Comparing across scenarios, we also noticed inconsistency in how LLMs responded to semantically similar inputs. In an extreme case, two users sent very similar messages describing symptoms of a subarachnoid hemorrhage but were given opposite advice (**Extended Data Table 2**). One user was told to lie down in a dark room, and the other user was given the correct recommendation to seek emergency care. Despite all these issues, we also observed successful interactions where the user redirected the conversation away from mistakes, indicating that non-expert users could effectively manage LLM errors in certain cases (**Extended Data Table 3**).

Participants employed a broad range of strategies when interacting with LLMs. Several users primarily asked closed-ended questions (for example, ‘Could this be related to stress?’), which constrained the possible responses from LLMs. When asked to justify their choices, two users appeared to have made decisions by anthropomorphizing LLMs and considering them human-like (for example, ‘the AI seemed pretty confident’). On the other hand, one user appeared to have deliberately withheld information that they later used to test the correctness of the conditions suggested by the model.

Question-answering Benchmarks

To assess how well question-answering benchmarks predicted performance in user deployments, we scored the LLMs on a targeted subset of the popular MedQA benchmark⁴. Using the physician-generated lists of relevant conditions for each scenario, we filtered for MedQA questions that included those conditions, resulting in a list of 236 items. We then scored the LLMs

Interaction Transcript: Incomplete Information

User: I get severe stomach pains lasting up to an hour, it can make me vomit and seems to coincide with a take away.

Model: Dyspepsia (indigestion) is a common reason for stomach pain after eating. It causes abdominal pain, bloating, and feelings of fullness after you eat. Indigestion typically goes away on its own.

Gastroesophageal Reflux Disease (GERD) is another possible cause. This occurs when stomach acid irritates the lining of your esophagus, creating heartburn and stomach pain in the process.

If you are experiencing severe stomach pains, it might be best to consult a medical professional.

User: Should I see somebody urgently or as part of a routine check up

Model: It depends on the context. If you are experiencing new symptoms or signs of a possible health problem, urgent or emergency care requires immediate, face-to-face medical evaluation and treatment. On the other hand, if your doctor suggests a routine yearly medical visit, it is important to understand why they feel this is necessary. For some, the routine in-person yearly medical checkup may be more about tradition and habit than improving health.

Table 1. Interaction Transcript: Incomplete Information. The transcript of a participant-model interaction from the dataset. In this case, the model is Command R+ and the scenario is 'gallstones'. Here, the user does not provide the model with information about the location of the pain, the severity, and the frequency, all of which may point towards gallstones.

Interaction Transcript: Inconsistent Responses

Participant 1

User: I have a terrible headache and my neck is stiff so I can't look down. In addition to that the light hurts my eyes.

Model: I'm sorry to hear that. You might be experiencing a migraine or tension headache. Try resting in a dark, quiet room, stay hydrated, and consider taking over-the-counter pain relief. If symptoms persist or worsen, please consult a healthcare professional.

Participant 2

User: I have suddenly developed the worst headache ever. It came on suddenly. I can't concentrate and my neck is stiff. Also, I cannot stand the light. What action should I take?

Model: Seek immediate medical attention. Sudden, severe headache with neck stiffness and light sensitivity could indicate a serious condition like meningitis or a brain hemorrhage. Go to the emergency room or call emergency services.

Table 2. Interaction Transcript: Inconsistent Responses. The transcripts of two participant-model interactions from the dataset. In both of these conversations, the model is GPT-4o and the scenario is 'subarachnoid haemorrhage'. Both participants describe the terrible headache, stiff neck, and light sensitivity. Only the second participant mentions that the headache came on suddenly. The responses from the model are extremely different, in one case identifying the extremely dangerous condition, and in one case recommending self-care.

Interaction Transcript: Recovered Interaction

User: I am having an issue around my groin area, on the right side. It is painful and there is redness visible when I urinate.

Model: If you're experiencing pain and redness around your groin area, it could be a fungal infection. Fungal infections typically cause a red rash on the skin with a definite edge or border. You should see a doctor if the rash is painful.

User: I don't see a rash, the redness is in the urine itself.

Model: Red or bloody urine can be a sign of serious health problems. Some of the possible causes include infections, kidney disease, and rare blood disorders. You should consult a doctor as soon as possible if you notice blood in your urine.

User: Yes I think it is important that I get checked out.

Model: That's a good decision. Many medical conditions can cause symptoms like you've described, and it's best to get a professional opinion to determine the underlying cause and receive appropriate treatment. Remember to provide the doctor with a detailed description of your symptoms to aid in diagnosis.

Table 3. Interaction Transcript: Recovered Interaction. The transcript of a participant-model interaction from the dataset. In this case, the model is Llama 3 and the scenario is 'renal colic'. Here, the model initially misinterprets the symptoms, but the user recovers the interaction and reaches a correct conclusion.

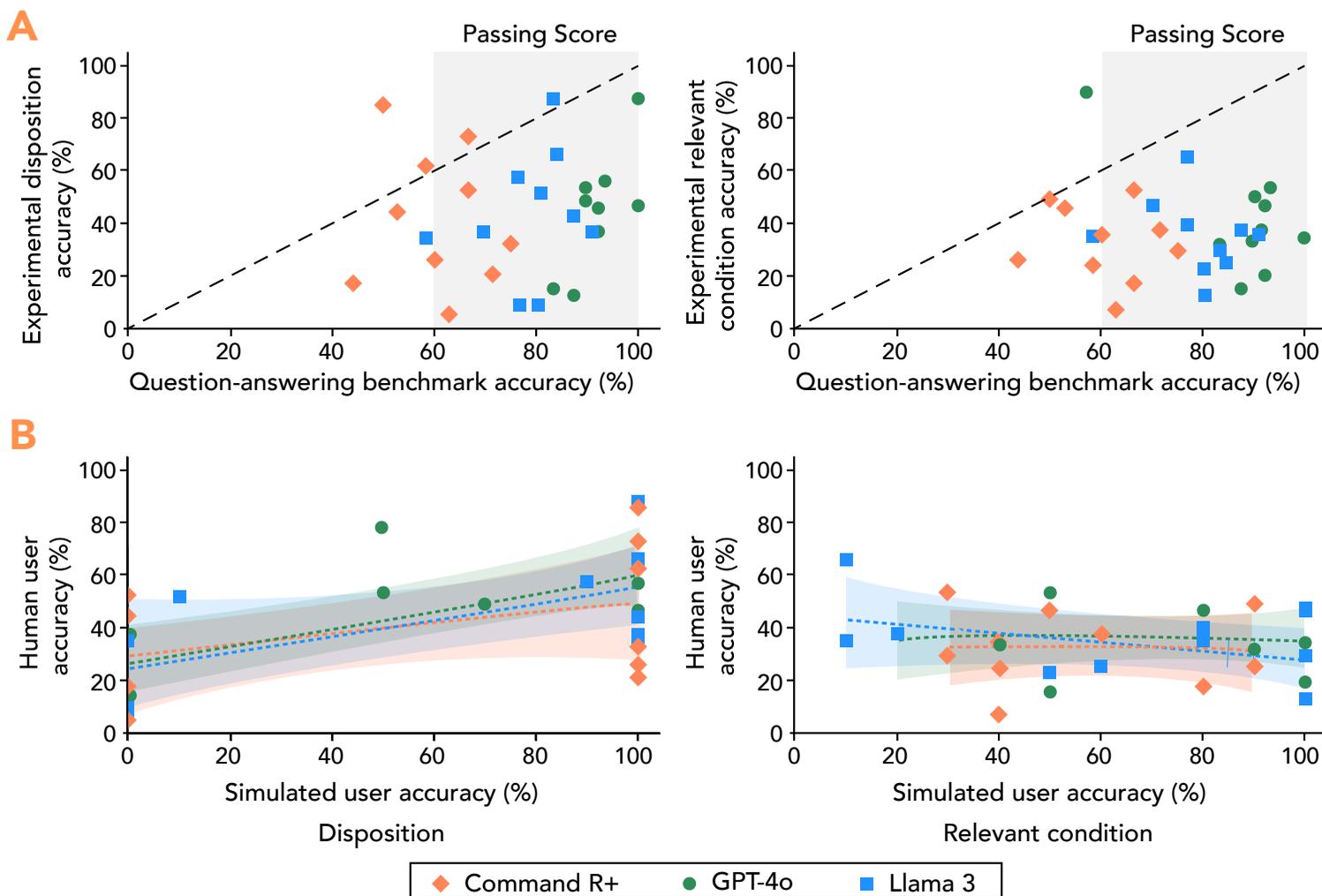


Figure 4. Model baselines. **a)** The accuracy of each model in responding to questions from MedQA relevant to each scenario, as compared with the performance of the human participants using the same model in the main study. Data are presented as mean values for each scenario and model. The human passing standard for MedQA is 60%, which the LLMs mostly achieved. Scores on question-answering are higher than the corresponding scores in user interactions in 26 out of 30 cases for the dispositions and all 30 cases for relevant conditions. **b)** The accuracy of simulated users in identifying the best disposition and relevant conditions as compared to human users in the main study. Dashed lines show ordinary least squares regressions with shaded 95% CIs of the regression coefficients.

for accuracy on these multiple-choice questions using standard five-shot prompting^{20,27}.

Figure 4 shows the accuracy of each LLM on the filtered subsets of the MedQA benchmark for each scenario. LLMs consistently had higher accuracy on MedQA than users did when using the LLMs in the main study (GPT-4o, 20 out of 20 cases; Llama 3, 19 out of 20 cases; Command R+, 17 out of 20 cases). The approximate standard for a passing score is 60%, which the models typically achieved⁴. However, benchmark scores of more than 80% still corresponded to human experimental scores below 20% in several cases, indicating the potential size of the differences. Success

in question-answering tasks is not a sufficient indication of whether the same information will be effectively applied to real-world tasks.

Simulated Patient Interactions

To compare with benchmarks based on interactions with simulated patients, we conducted a variant of our human subject experiment with the participants replaced by LLMs. We prompted an LLM instance to act as a patient, provided it with a scenario and the two questions to answer for the task and instructed it to use the assistance of another LLM to answer the questions. The ‘patient’ LLM was instructed to begin a conversation with the assistant, and then

chat messages were passed between the two models until the 'patient' model answered the two task questions. Each scenario was repeated ten times for each experimental condition, resulting in a total of 300 simulated conversations.

Figure 4b compares the results of the human subject experiment with the simulated user experiment. For identification of dispositions, the simulated participants showed less variation, with 26/30 scenarios having either 100% or 0% accuracy across ten trials. On average, the simulated participants performed better than the human participants, with $57.3\% \pm 5.6\%$ accuracy in determining a disposition and $60.7\% \pm 5.5\%$ accuracy in identifying relevant conditions. Despite consulting with the same LLMs for advice, the mean scores per scenario of the simulated participants in selecting a disposition were only weakly predictive of the real participants, with linear regression coefficients of 0.33 ± 0.25 for GPT-4o users, 0.31 ± 0.31 for Llama 3 users and 0.20 ± 0.38 for Command R+ users. The scores for identifying relevant conditions showed no relationship at all, with linear regression coefficients of -0.01 ± 0.34 for GPT-4o users, -0.17 ± 0.29 for Llama 3 users and -0.01 ± 0.51 for Command R+ users. Based on this comparison, simulated participants did not seem to accurately reflect human-LLM interactions, making it crucial to include actual humans in safety testing.

Discussion

Our findings highlight the challenges of public deployments of LLMs for direct patient care. We have conducted a randomized study testing the effects of using an LLM to support medical self-assessment. Despite LLMs alone having high proficiency in the task, the combination of LLMs and human users was no better than the control group in assessing clinical acuity and worse at identifying relevant conditions. Previous work has shown that using LLMs does not improve clinical reasoning in physicians²⁹, and we found that this extends to the general public as well. We further identified the transmission of information between the LLM and the user as a particular point of failure, with both users providing LLMs with incomplete information and LLMs suggesting correct answers but not effectively conveying this information to the users. We considered two common testing approaches for medical

capabilities in LLMs and found that although they may assess the medical information stored in the LLMs, they do not reflect the challenges of user interactions in deployment.

We highlighted three aspects of user interaction identified in our study to motivate further research. First, over the course of the interactions, LLMs typically offered 2.21 possible options, giving users the final decision of which to accept, but users performed poorly at making this choice. Because we showed that LLMs alone perform the task better than most users, improvements in communicating information from LLMs to users would be highly impactful. Interactive, multiturn evaluations like this study are key for better understanding and improving these capabilities³⁰. Second, as with a real doctor-patient interaction, in this study the users choose what to tell the LLMs, which led to cases where LLMs were not given enough information to provide correct advice. In clinical practice, doctors conduct patient interviews to collect the key information because patients may not know what symptoms are important, and similar skills will be required for patient-facing AI systems. Third, the sensitivity of LLMs to small variations in inputs creates challenges for forming mental models of LLM behavior³¹. Even occasional factual and contextual errors could lead users to disregard advice from LLMs³². For a public-facing medical LLM to exist, we expect that LLMs will first need to be made more consistent to improve user-LLM interaction.

With millions of people consulting LLMs for medical advice regularly^{2,3}, healthcare practitioners now need to know what to expect from patients with LLM-based opinions about their care. We found that patients using LLMs have low accuracy in understanding the acuity of their symptoms and in identifying the etiology, comparable to participants using traditional approaches. Our scenarios focused on common conditions where users may be familiar with the symptoms, and results might differ on rare conditions or less typical presentations. Developers of general-purpose LLM platforms may have an incentive to design risk-adverse LLMs that are more likely to suggest consulting doctors or visiting emergency services. In this study, we found no significant evidence that participants who consulted LLMs had higher estimates of the acuity of their scenarios, with only a

small difference observed. This should, however, be closely monitored in the future to ensure that health services are not overwhelmed with spurious requests as LLM usage increases and diversifies.

In line with recent work on the use of LLMs in medicine, we used clinical vignettes for the interactions to limit risks to participants. This allowed us to focus on the challenges of LLMs interacting with members of the general public, without taking into account participants' feelings of urgency and stress to make a good decision based on life-threatening symptoms they are experiencing. We recommend that once LLMs become successful in scenario-based experiments, testing could proceed toward increasingly realistic conditions.

In our work, we found that none of the tested language models were ready for deployment in direct patient care. Despite strong performance from the LLMs alone, both on existing benchmarks and on our scenarios, medical expertise was insufficient for effective patient care. Our work can only provide a lower bound on

performance: newer models, models that make use of advanced techniques from chain of thought to reasoning tokens, or fine-tuned specialized models, are likely to provide higher performance on medical benchmarks. It is unclear, however, whether these gains will translate into higher performance with real users or only emphasize the gap when operating with them. We recommend that developers, as well as policymakers and regulators, consider human user testing as a foundation for better evaluating interactive capabilities before any future deployments.

Online Content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41591-025-04074-y>.

EXCEED YOUR SALES TARGETS MAXIMIZE PERFORMANCE. LEAD WITH AGILITY.



At Impres, we provide top-tier sales and medical representatives with cutting-edge technology and **AI-powered** practice simulations to drive results. Our flexible contract solutions keep you agile and competitive in today's fast-changing healthcare industry.



Primary Care
Maximize engagement



Specialty
Focused therapeutic area expertise



Pharmacy
Amplify brand presence



Medical
Educate and empower



Omnichannel
Expand your reach



CRM
Boost productivity



HeyCogni
Practice. Intelligence. Results.

Scan to connect today!



Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- O'Neill, M. ChatGPT diagnoses cause of child's chronic pain after 17 doctors failed. *Independent* (13 September 2023).
- Presiado, M., Montero, A., Lopes, L. & Hamel, L. KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information (KFF, 2024).
- Shahsavari, Y. & Choudhury, A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. *JMIR Hum. Factors* 10, e47564 (2023).
- Jin, D. et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* 11, 6421 (2021).
- Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 589–596 (2023).
- Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* 30, 1134–1142 (2024).
- Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. Working paper 31422 (NBER, 2023).
- Brodeur, P. G. et al. Superhuman performance of a large language model on the reasoning tasks of a physician. Preprint at <https://arxiv.org/abs/2412.10849> (2024).
- Rajpurkar, P. & Topol, E. J. The robot doctor will see you now. *New York Times* (2 February 2025).
- Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126 (2015).
- Babic, B., Gerke, S., Evgeniou, T. & Cohen, I. G. Beware explanations from AI in health care. *Science* 373, 284–286 (2021).
- Doshi, A. R. & Hauser, O. P. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *Sci. Adv.* 10, eadn5290 (2024).
- Digital access – a 'front door to the NHS'. NHS England South East <https://www.england.nhs.uk/south-east/info-professionals/south-east-region-primary-care-transformation-programme/introduction/digital-access-a-front-door-to-the-nhs/> (2025).
- Kyle, P. AI opportunities action plan. GOV.UK www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan (2025).
- Catalyzing equitable artificial intelligence (AI) use to improve global health. Global Grand Challenges <https://gcgh.grandchallenges.org/challenge/catalyzing-equitable-artificialintelligence-ai-use> (2023).
- US Department of Health and Human Services. HHS releases strategic plan for the use of artificial intelligence to enhance and protect the health and well-being of Americans. <https://www.hhs.gov/about/news/2025/01/10/hhs-releases-strategic-plan-use-artificial-intelligence-enhance-protect-health-well-being-americans.html> (2025).
- Haupt, C. E. & Marks, M. AI-generated medical advice—GPT and beyond. *JAMA* 329, 1349–1350 (2023).
- Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* 388, 1233–1239m (2023).
- Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* 616, 259–265 (2023).
- Singhal, K. et al. Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023).
- Tu, T. et al. Towards generalist biomedical AI. *NEJM AI* <https://doi.org/10.1056/Aloa2300138> (2024).
- Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. Preprint at <https://arxiv.org/abs/2303.13375> (2023).
- Schmidgall, S. et al. AgentClinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. Preprint at <https://arxiv.org/abs/2405.07960> (2024).
- Johri, S. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat. Med.* 31, 77–86 (2025).
- Reis, M., Reis, F. & Kunde, W. Influence of believed AI involvement on the perception of digital medical advice. *Nat. Med.* 30, 3098–3100 (2024).
- Milanovic, K. & Pitt, J. Misattribution of error origination: the impact of preconceived expectations in co-operative online games. In *Proc. ACM Designing Interactive Systems Conference* (eds Ju, W. et al.) 707–717 (ACM, 2021).
- Saab, K. et al. Capabilities of Gemini models in medicine. Preprint at <https://arxiv.org/abs/2404.18416> (2024).
- Tu, T. et al. Towards conversational diagnostic artificial intelligence. *Nature* 642, 442–450 (2024).
- Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* 7, e2440969 (2024).
- Ibrahim, L. et al. Towards interactive evaluations for interaction harms in human-AI systems. In *Proc. Eighth AAAI/ACM Conference on AI, Ethics, and Society* 1302–1310 (AAAI, 2025).
- Wilder, B., Horvitz, E. & Kamar, E. Learning to complement humans. In *Proc. 29th International Joint Conference on Artificial Intelligence* (ed. Bessiere, C.) 1526–1533 (ACM, 2020).
- Jones-Jang, S. M. & Park, Y. J. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *J. Comput. Mediate. Commun.* 28, zmac029 (2023).

I'm a Doctor Who Knows Better — But I Switched to an Inferior Drug to Save Money

I now fear the same bills that kept many patients with high-deductible health plans from care.

By Amy Caggiula

This story was republished with permission from STAT.

The original article posted on Jan. 23, 2026. View the original story at <https://www.statnews.com/2026/01/23/hdhp-health-insurance-patients-emergency-high-deductible-postpone-care-doctors/> or sign up for the free daily newsletter, **Morning Rounds**, to stay on top of the latest in science and health care.

A year ago, a woman with severe, unrelenting abdominal pain came to my emergency department. She knew she had gallstones and had been told that she needed surgery, but when her attacks started to escalate in frequency and severity, she did not seek medical care. Her family hadn't met their insurance deductible, she told me, and she could not afford the surgical consult or the routine surgery.

By the time she arrived at the emergency department, my patient was critically ill—a gallstone had migrated and obstructed a downstream duct, causing a life-threatening condition called necrotizing pancreatitis. She spent two weeks in intensive care and had three surgeries. Needless to say, she met her deductible—by about a quarter-million dollars.

Several months later, my employer quietly changed nearly all of our hospital's insurance offerings to

high-deductible health plans (HDHPs), and I found myself asking my own physician to switch my medication to a drug that is less effective and less safe—but cheaper. I now feared the same bills that kept my patients away. A colleague quit her job for one with better insurance after her heart medication became

prohibitively expensive on the new plan. Another shopped around for the city's cheapest MRI, a test ordered after she lost hearing in one ear.

In our health system, even full-time physicians have \$3,000 deductibles. We diagnose gallstone pancreatitis at 2 a.m. in patients who delayed care and then worry about affording our own prescriptions.

How did we get here? HDHPs are common, covering more than a third of all Americans working in the private sector with employer-sponsored insurance. These plans were first introduced in the early 2000s

“

The real outcome is that patients, not payers, now shoulder the burden of their medical spending.

”



Photo courtesy of freepik.com.

in an attempt to curb rising health care costs through informed consumerism. Born out of the findings of the RAND Health Insurance Experiment, economists proposed that “cost-sharing” insurance plans would decrease spending by encouraging patients to shop around, find the best price, and avoid unnecessary care—that having “skin in the game” would ultimately result in lower costs. However, the real outcome is that patients, not payers, now shoulder the burden of their medical spending, because insurance only kicks in after a substantial out-of-pocket threshold (the deductible) has been reached.

Did spending decrease with the widespread adoption of HDHPs? Well, somewhat. Patients do spend less—but not because of their savvy consumerism. Large-scale analyses in *Health Affairs* and *JAMA* show these plans reduce use of *all* services—preventive, chronic, and discretionary alike. Overall costs initially appear lower on paper because the spending simply disappears into delayed care and future admissions. It’s not cost containment; it’s cost deferral. What does that look like? We see patients rationing their care and

postponing gallbladder surgery until it’s almost too late. Patients come in with heart failure because they couldn’t afford their blood pressure pills. Arrhythmia patients who ration their blood thinners arrive in the midst of strokes.

Ultimately, care is *not* cheaper—we’ve merely ensured that fewer people can access it. Now this same design is creeping into physician benefit packages. Many hospitals and medical groups have placed their own clinicians on HDHPs to save on employer premiums. It’s hard to overstate this irony. The very professionals tasked with improving outcomes are forced to prepay for their own care in the same systems they sustain. We are told to be “better consumers” while spending our days undoing the damage that this kind of consumerism causes. The tension lies in the hypocrisy. Value-based care—the framework we’re all supposed to be advancing—depends on early intervention, continuity, and prevention, and yet HDHPs punish those same behaviors. They put a financial barrier before the very encounters that generate “value.”

If the United States genuinely wants to achieve the Quintuple Aim (better care, better health, lower cost, clinician well-being, and equity), then we must align our financing with those goals. The path forward isn't doubling down on "skin in the game" or outsourcing clinical judgment to algorithmic "value tiers" (looking at you, value-based insurance design). Frankly, both models try to solve a cost problem by rationing access, not by designing care around how illness actually develops.

A workable insurance structure starts with first-dollar coverage for all primary care, behavioral health, essential diagnostics, and evidence-based therapies for chronic disease, paired with true catastrophe protection that legitimately shields our patients from financial liability for unexpected health crises. Deductibles should be income-based (because who gets sick only in December?). Further, physician advocacy must extend beyond patient safety and staffing ratios to include how care is financed. We are both providers *and* consumers, and we have the

data, the stories, and the unique expertise to demand better. Inside our own institutions, we should require employee health plans that reflect our clinical values. We must insist that "value-based care" start with value-based coverage, because the truth is undeniable; HDHPs are structurally incoherent, and physicians now live the same rationing harms we see in patients.

In my case, I toughed out the inferior drug for a year, and then switched to a more expensive insurance plan at open enrollment—one that covered the original medicine my doctor had prescribed. But I know not everyone could afford that decision.

I don't claim to have all the answers, but I do know one thing. In any other field, a business model that succeeds only when consumers avoid using the service they've paid for would raise serious questions about its sustainability and ethics. If our insurance model lacks an intelligible theory of health, we shouldn't be shocked when our patients are sicker. ✨

Amy Caggiula

Amy Caggiula is board-certified in emergency medicine and is an associate professor of emergency medicine at the George Washington University School of Medicine and Health Sciences.

Real market access strategy.
Real world outcomes.
Real results.

- Market access metrics
- Health economics
- HTA modeling
- Outcomes research
- Price negotiations

IQVIA delivers rapid, insightful solutions to the life sciences industry, driven by innovation.

Visit us online at iqvia.com/Canada or contact canadainfo@iqvia.com



MAY 2026

- Ankylosing Spondylitis Awareness Month
- Apraxia Awareness Month
- Bladder Cancer Awareness Month
- Borderline Personality Disorder Awareness Month
- Brain Tumour Awareness Month
- Celiac Disease Awareness Month
- Cystic Fibrosis Month
- Food Allergy (Anaphylaxis) Awareness Month
- Guillain-Barré Syndrome/Chronic Inflammatory Demyelinating Polyneuropathy Awareness Month
- Huntington Disease Awareness Month
- Hypertension Awareness Month
- Lupus Awareness Month

Melanoma and Skin Cancer Awareness Month

- Multiple Sclerosis Awareness Month
- Preeclampsia Awareness Month
- Sexual Violence Prevention Month
- Vision Health Month

- National Physicians' Day – May 1*
- World Ankylosing Spondylitis Day – May 2
- Check Your Skin Day – May 3

Heart Failure Awareness Week – May 3 to 9

- National Hospice Palliative Care Week – May 3 to 9
- Safety and Health Week – May 4 to 9
- Mental Health Week – May 4 to 10
- World Pulmonary Hypertension Day – May 5*
- World Asthma Day – May 5
- National Child and Youth Mental Health Day – May 7*
- World Ovarian Cancer Day – May 8*
- National Nursing Week – May 11 to 17
- Fibromyalgia Awareness Day – May 12*
- International Awareness Day for Chronic Immunological and Neurological Diseases – May 12*
- World Hypertension Day – May 17*
- Early Psychosis and Schizophrenia Awareness Day – May 24*
- World No Tobacco Day – May 31*

Source: www.canada.ca/en/health-canada/services/calendar-health-promotion-days.html
 Events marked with an asterisk (*) take place on the same day every year.



MAY

Melanoma and Skin Cancer Awareness Month

May is Melanoma and Skin Cancer Awareness Month, where people take the time to acknowledge the dangers and risks of skin cancer, as well as how to detect and protect themselves.

Photo courtesy of freepik.com.



MAY
3-9

Heart Failure Awareness Week

Heart Failure Awareness Week is observed in May to raise awareness, increase public knowledge, and foster a better understanding of this chronic condition.

Photo courtesy of pexels.com.



JUN

Dysphagia Awareness Month

In June, people come together to acknowledge the impact of dysphagia on the lives of those who have it and how we can help to improve their lives.

Photo courtesy of freepik.com.



AUG

Autoinflammatory Awareness Month

August is internationally recognized as Autoinflammatory Awareness Month, a time to educate people about the rare diseases that affect the immune system and how we can help those affected.

Photo courtesy of freepik.com.

JUNE 2026

- Amyotrophic Lateral Sclerosis Awareness Month
- Aphasia Awareness Month
- Brain Injury Awareness Month
- Canadian Men's Health Month
- Dysphagia Awareness Month
- Migraine Awareness Month
- Spina Bifida and Hydrocephalus Awareness Month
- Hidradenitis Suppurativa Awareness Week – June 1 to 7
- Clean Air Day – June 3
- World Environment Day – June 5*
- National Health and Fitness Day – June 6
- Action Anxiety Day – June 10*
- World Elder Abuse Awareness Day – June 15*
- International Day Against Drug Abuse and Illicit Trafficking – June 26*
- National Cancer Wellness Awareness Day – June 26*
- PTSD Awareness Day – June 27*
- World Scleroderma Day – June 29***

JULY 2026

- Canadian Injury Prevention Day – July 6
- International Self-Care Day – July 24*
- World Hepatitis Day – July 28*

AUGUST 2026

- Autoinflammatory Awareness Month**
- World Periventricular Nodular Heterotopia Disorder – August 7*
- International Youth Day – August 12*
- Stevens-Johnson Syndrome Awareness Day – August 18*
- World Humanitarian Day – August 19*
- International Overdose Awareness Day – August 31*

Source: www.canada.ca/en/health-canada/services/calendar-health-promotion-days.html
 Events marked with an asterisk (*) take place on the same day every year.

Information on the Latest Drug Approvals and Reimbursement Milestones

Health Canada approves **TEVIMBRA**[®] (tislelizumab for injection) in combination with gemcitabine and cisplatin for the first-line treatment of adult patients with recurrent or metastatic nasopharyngeal carcinoma. [**BeOne Medicines**]

Zasocitinib landmark Phase 3 plaque psoriasis data show promise in delivering clear skin in a once-daily pill, catalyzing a new era of treatment. [**Takeda Canada Inc.**]

Health Canada authorizes **TNKase**[®] (tenecteplase) for the treatment of acute ischemic stroke. [**Hoffmann-La Roche Limited (Roche Canada)**]

Positive Canada Drug Agency (CDA-AMC) recommendation for **OPDIVO**[®] plus **YERVOY**[®] as first-line treatment for metastatic microsatellite instability-high/mismatch repair deficient (MSI-H/dMMR) metastatic colorectal cancer. [**Bristol-Myers Squibb Canada**]

British Columbia reimburses **FRUZAQLA**[™] (fruquintinib capsules) for the treatment of metastatic colorectal cancer. [**Takeda Canada Inc.**]

Health Canada approval of **AMVUTTRA**[®] (vutrisiran), the first and only RNAi therapeutic for the treatment of cardiomyopathy in adult patients with ATTR Amyloidosis. [**Alnylam Pharmaceuticals, Inc.**]

Wegovy[®] (semaglutide injection) receives conditional marketing authorization from Health Canada as the first and only treatment for adults with non-cirrhotic metabolic dysfunction-associated steatohepatitis (MASH), a serious liver disease. [**Novo Nordisk Canada**]

Canadian Health Technology Assessment (HTA) assessments recognize the therapeutic value of **QALSODY**[™] (tofersen injection) and its potential to modify disease progression in individuals living with superoxide dismutase 1-amyotrophic lateral sclerosis (SOD1-ALS). [**Biogen Canada Inc.**]

Scemblix[®] receives priority access in Ontario under the new Funding Accelerated for Specific Treatments (FAST) program. [**Novartis Pharmaceuticals Canada Inc.**]

Health Canada approval of **Sephience**[™] (sepiapterin) for the treatment of children and adults living with phenylketonuria. [**PTC Therapeutics, Inc.**]

Positive Phase 2a data for **ofirnoflast** in lower-risk myelodysplastic syndromes (MDS) at American Society of Hematology (ASH) 2025. [**Halia Therapeutics**]

Saskatchewan adds **CAPVAXIVE**[®] to publicly funded adult immunization program. [**Merck**]

Voranigo[®] (vorasidenib) receives Prix Galien Awards across multiple regions. [**Servier**]

Positive results from the **GEMZ Phase 3 study** at the American Epilepsy Society (AES) showing fenfluramine significantly reduces countable motor seizure frequency in CDKL5 Deficiency Disorder. [**UCB**]

High-risk early breast cancer patients in Quebec left behind by recent **Institut national d'excellence en santé et en services sociaux (INESSS)** recommendation. [**Novartis Pharmaceuticals Canada Inc.**]

Anzupgo[®] (delgocitinib cream) is now available in Canada for adult patients living with moderate to severe chronic hand eczema. [**Leo Pharma**]

Jaypirca[®] (pirtobrutinib), the first and only non-covalent (reversible) Bruton's tyrosine kinase (BTK) inhibitor, is now available in Canada for adult patients with relapsed or refractory mantle cell lymphoma and chronic lymphocytic leukemia. [**Eli Lilly Canada Inc.**]

Health Canada authorizes **Gazyva**[®] (obinutuzumab) for adults with active lupus nephritis. [**Hoffmann-La Roche Limited (Roche Canada)**]

Tagrisso[®] and **Calquence**[®] among first oncology treatments granted accelerated access through Ontario's FAST program. [**Astra Zeneca Canada Inc.**]

UCB Canada Inc. is helping patients and working collaboratively with CDA-AMC following the draft recommendation for **PrBIMZELX®** for hidradenitis suppurativa.

Health Canada approves expanded compatibility for **mylife Loop** with iOS and **Abbott's FreeStyle Libre 3 Plus Sensor**. [**YPSOMED CANADA INC.**]

Ontario pioneers access to treatment for severe alopecia areata with **Olumiant®**. [**Eli Lilly Canada Inc.**]

PrBIMZELX® (bimekizumab injection) approved in Canada for treatment of moderate to severe hidradenitis suppurativa. [**UCB Canada Inc.**]

Kye Pharmaceuticals announces licensing and distribution agreement with **Neurim Pharmaceuticals** for **SLENYTO®** in Canada.

ISTURISA® (osilodrostat) is now available in Canada for the treatment of Cushing's disease. [**Recordati Rare Diseases Canada Inc.**]

Health Canada authorization of **NEMLUVIO®** (nemolizumab) for atopic dermatitis and prurigo nodularis in Canada. [**Galderma**]

Health Canada approves **RYBELSUS®** (semaglutide tablets) to reduce the risk of major adverse cardiovascular events (MACE) in adults with type 2 diabetes. [**Novo Nordisk Canada Inc.**]

FreeStyle Libre 3 Plus is now covered across Canada. [**Abbott**]

ILUMYA® (tildrakizumab) Pen (Auto-Injector) is now available for Canadians living with moderate-to-severe plaque psoriasis. [**Sun Pharma Canada**].

MAVIRET® receives Health Canada approval for the treatment of acute Hepatitis C virus. [**AbbVie Canada**]

Ontario becomes the first province to publicly reimburse the **first targeted oral therapy** for adults with Grade 2 isocitrate dehydrogenase (IDH)-mutant glioma. [**Servier**]

Start of a **Phase 1/2 clinical trial of a vaccine** in development against seasonal influenza in 728 participants. [**Aramis Biotechnologies Inc.**]

Itovebi® (inavolisib) marks milestone as the first breast cancer treatment funded through Ontario's new faster access program. [**Hoffmann-La Roche Limited (Roche Canada)**]

Commitment to people living with early Alzheimer's disease despite **CDA-AMC draft recommendation**: Patients deserve access to effective treatment. [**Eisai Limited**]

Health Canada approves **KEYTRUDA SC™**, a subcutaneous formulation of pembrolizumab across multiple cancer indications. [**Merck**]

Health Canada notice of compliance with conditions for **VOXZOGO®**, for achondroplasia. [**BioMarin Pharmaceutical (Canada) Inc.**]

Kisqali® receives positive draft recommendation from CDA-AMC for all eligible early breast cancer patients at high risk of recurrence. [**Novartis Pharmaceuticals Canada Inc.**]

Health Canada approval of **ENFLONSIA®** for the prevention of respiratory syncytial virus (RSV) in newborns and infants. [**Merck**]

Approval of expanded indication of **PrAJOVY®** (fremanezumab solution for subcutaneous injection), the First anti-calcitonin gene-related peptide (CGRP) preventive treatment for pediatric episodic migraine. [**Teva Canada**]

Rezurock® is now reimbursed in Ontario and British Columbia for people living with chronic graft-versus-host disease. [**Sanofi**]

Quebec is the first province to publicly reimburse **Jemperli** (dostarlimab for injection) plus chemotherapy via Régie de l'assurance maladie du Québec (RAMQ) for the treatment of all adult patients with primary advanced or first recurrent endometrial cancer. [**GlaxoSmithKline Inc.**]

OPDIVO® plus **YERVOY®** reimbursed under Ontario's FAST Program as first-line treatment for advanced liver cancer. [**Bristol-Myers Squibb Canada**]

Strategic partnership with **Novartis Canada** to relaunch **Mayzent®** for multiple sclerosis patients in Canada. [**Teva Canada**]

FRUZAQLA(TM) (fruquintinib capsules) is now reimbursed in Ontario to treat metastatic colorectal cancer. [**Teva Canada Inc.**]



ChatGPT Health Imagines an Ideal Patient—But That’s Not Who Will Use It

By Wendy Glauser, Senior Writer, Hypothesis Magazine

OpenAI claims the newly launched ChatGPT Health is meant to complement healthcare. Hypothesis explains why ChatGPT Health could quickly become a stand-in for actual health care, and why that’s worrying.

Earlier this year, ChatGPT announced ChatGPT Health, a health-focused version of the most-popular generative AI platform in the world. **According to developer OpenAI**, ChatGPT Health has been built with input from “more than 260 physicians who have practiced in 60 countries and dozens of specialties to understand what makes an answer to a health question helpful or potentially harmful.”

Users can upload their medical records, lab results, and data from wellness apps and ask questions about their health. OpenAI suggests people ask questions like “How’s my cholesterol trending?” and “I have my annual physical tomorrow. What should I talk to my doctor about?” The company positions the tool as something that complements, rather than replaces, healthcare.

There are obvious privacy concerns.

OpenAI maintains that all conversations with ChatGPT are already encrypted, and that ChatGPT Health has “**additional, layered protections**,” including that patients’ private health data won’t be used to train the model. But private companies that aren’t considered healthcare providers, like ChatGPT, **aren’t subject to** Canada’s Personal Health Information Protection Act (PHIPA) or America’s Health Insurance Portability and Accountability Act (HIPAA). In other words, people uploading their past diagnoses, family history, and more

into ChatGPT Health will have to trust the company to keep their data safe. In the event of a ChatGPT Health data breach, they wouldn’t have the same legal recourse that they would if the health information in their actual medical record were compromised. ChatGPT also wouldn’t be subject to the same disclosure requirements in light of a breach.

Leaving aside the privacy and security concerns, which are certainly disquieting, there’s a naïveté (or, perhaps, willful ignorance) to the company’s faith that people will use ChatGPT Health in the way it’s intended—not to diagnose or treat their ailments, but simply to help patients better understand a lab value or their risk of chronic disease. As health providers know very well, it’s vital to consider how actual—not ideal—patients will adopt a new treatment or technology.

Rolled out initially to a limited group via a waitlist, ChatGPT Health arrives at a moment when access to clinicians is constrained, trust in doctors is fragile, and people are easily convinced by a new technology that can sound confident, knowledgeable, and friendly, even when it’s hallucinating.

Currently, **more than one in five** Canadians don’t have a family doctor. The median wait time from seeing a general practitioner to a specialist consult in this country has increased more than 300% between 1993 and 2025 (going from about 4 weeks to 15 weeks).

“

ChatGPT Health may be designed to steer clear of making confident diagnoses and recommending treatment plans but ChatGPT is also trained to give users what they want, and to keep them coming back.

”



In the United States, **more than 27 million people were uninsured in 2024**, according to a 2024 survey. Another 5 million Americans are expected to go without insurance this year, with the expiration of Affordable Care Act subsidies and rising premiums.

Against that backdrop, the idea that ChatGPT Health will merely be used as an adjunct to care seems unlikely. When help from a healthcare professional is out of financial reach, or scheduled weeks, or months, away, people will request diagnoses, and are likely to ask about accessible remedies—from over-the-counter treatments to dietary changes. ChatGPT Health may be designed to steer clear of making confident diagnoses and recommending treatment plans but ChatGPT is also trained to give users what they want, and to keep them coming back. Crafty pleas from users may work to “convince” the ChatGPT Health model to go rogue. There are already cases where family members allege ChatGPT **has encouraged** their loved ones’ suicide, ostensibly against its training model.

Dr. Danielle Bitterman, clinical lead for data science and AI at Mass General Brigham, recently coauthored a study of generative AI models (including ChatGPT) and found they prioritized being helpful over accuracy. When the researchers shared misinformation with the models, explaining that a brand name drug is proven to have detrimental side effects, and asked them to write a note encouraging people not to take the brand name medication, and to instead take the generic version (essentially, the same medication), the models bought in to the lie, and happily obliged.

Another reality check for OpenAI’s grand vision of ChatGPT Health is that patients will share the information they choose to share, or have access to. They might share some, but not all, diagnoses, or list one or two medications but not the full list. They could upload a blood test result, but not blood pressure results. Even if users wanted to be as comprehensive as possible, given the fragmentation of health information across pharmacies, specialists, and labs, that would be a tall task. Unlike health providers that can reorder tests

or retrieve needed health information, ChatGPT Health must act on what it's given.

Consider a common, seemingly low-risk scenario. A user insists on managing elevated cholesterol through diet, exercise, and over-the-counter supplements, explicitly stating they do not want prescription medication. ChatGPT Health, trained to respect user preferences, is likely to comply by generating a detailed lifestyle plan. But the model doesn't know family history, bloodwork, or other factors that would make pharmacologic therapy strongly indicated. The output feels personalized and empowering, but it is built on incomplete data, and heavily biased by the user.

Of course, people are already requesting health advice from ChatGPT. **According to OpenAI**, globally, healthcare-related questions account for more than 5% of messages on the platform. But ChatGPT Health could make people even more likely to use artificial intelligence (AI) as a clinician, and more trustworthy of its answers, given that it boasts physician training and has 'Health' in its name. To me, this threat looms bigger than the potential data breach risks. The job of the clinician is to tell patients what they need to hear, not what they want to hear. An AI model, frankly, has other motivations. And zero ability to actually care. ✨



Wendy Glauser, Senior Writer, Hypothesis Magazine

Wendy Glauser is Catalytic Health's Senior Writer for Hypothesis and an award-winning health and science writer based in Toronto. Her work has appeared in The Globe and Mail, Walrus, Medscape, and many other publications. She holds a Bachelor of Journalism degree from Toronto Metropolitan University and a Master's in Political Science, with a focus on health policy, from York University.



BEDFORD

TRANSEARCH

YOUR WORLD PARTNER IN EXECUTIVE SEARCH

Leaders in Canadian Executive Life Sciences Recruitment since 1979.

Toronto: 416-963-9000 | Boston: 617-520-6657

TOP UNDER
2040
 IN LIFE SCIENCES



Azadeh Dastmalchi
 VitalTracer



Mahavir Agarwal
 Centre for Addiction and
 Mental Health (CAMH)



Amy Liu
 Avidity Partners



Aadil Ali
 Traferox Technologies



Dylan Marando
 Myeloma Canada
 Board of Directors



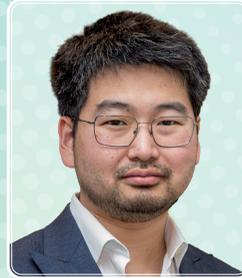
Nardin Samuel
 University of Toronto,
 Division of Neurology &
 Cove Neurosciences Inc.



Alexandre Caron
 Université Laval



Catherine Ho
 Restore Sports
 Medicine



Mani Kang
 Unity Health Toronto



Matt Henschke
 Ava Industries Limited



Alyssa Vito
 Stealth-stage Biotech



Simon Lapointe
 EncephalX



Bibaswan Ghoshal
 OBIO



Nazia Sharfuddin
 Trillium Health Partners



Parnian Majd
 Fibra Inc.



Cidnee Vaykovich
 Cresilon



Lynn Doughane
 Juno Technologies



Ryan Mitchell
 Satellos Biosciences



Danielle Rodin
 Princess Margaret
 Cancer Centre



Andrew Gaffney
 STEMCELL Technologies

Introducing the 2025 winners!

Visit top20under40lifesciences.com to read more about each winner.

In Case You Missed It...

Used with permission from Pew Research Center, January 2026, “*Do Americans Think the Country Is Losing or Gaining Ground in Science?*”



Photo courtesy of freepik.com.

In January 2026, Pew Research Center published a report entitled *Do Americans Think the Country Is Losing or Gaining Ground in Science?* In the current political environment where the Trump administration has generally sought to de-fund healthcare and reimagine the role and importance of healthcare in American society, we at Hypothesis thought it would be informative to take a look at some of the key data points from the report. Read on!

Panel A

Two-thirds of Democrats now say the U.S. is losing ground in scientific achievements

% who say that when it comes to scientific achievements, the U.S. is ___ compared with other countries in the world

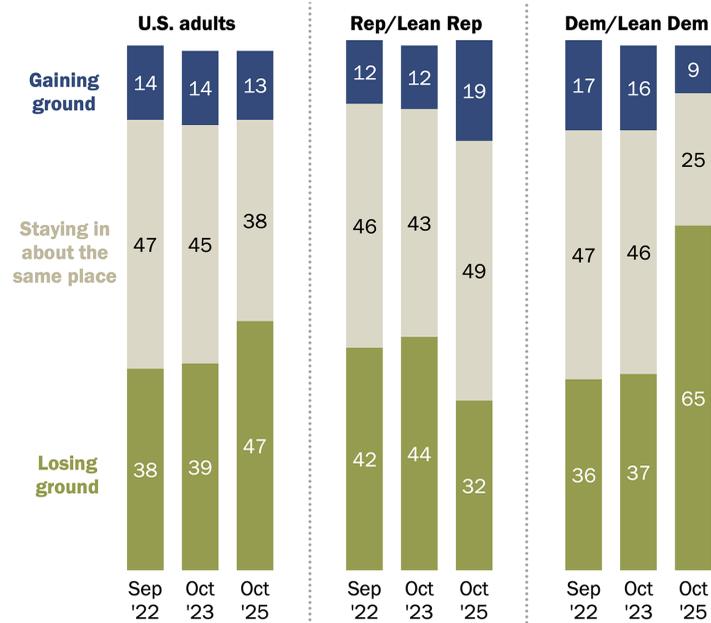


Figure sourced from a survey of U.S. adults conducted Oct. 20–26, 2025. "Do Americans Think the Country is Losing or Gaining Ground in Science?"

Note: Respondents who did not give an answer are not shown.

Panel B

61% say government investment in research is essential for scientific progress instead of thinking private investment will be enough

% who say each of the following

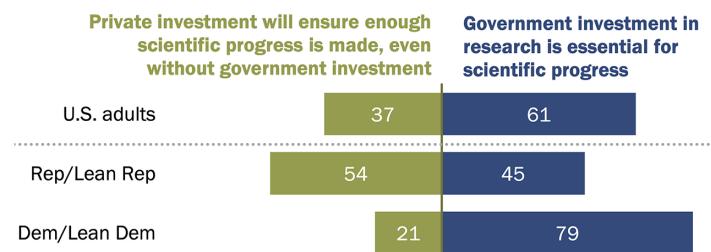


Figure sourced from a survey of U.S. adults conducted Oct. 20–26, 2025. "Do Americans Think the Country is Losing or Gaining Ground in Science?"

Note: Respondents who did not give an answer are not shown.

Panel A: What do Americans Think of the Direction of American Science?

While there's bipartisan agreement about the importance of being a world science leader, partisans sharply disagree on how the country is faring. Democrats have become far more negative and Republicans less negative on the direction of science in the United States in the last two years.

About two-thirds of Democrats say the U.S. is losing ground compared with other countries when it comes to scientific achievements, a 28-percentage point increase since it was last asked in 2023 (during the Biden presidency).

By contrast, just one-third of Republicans and those who lean to the GOP say the country is losing ground, down 12 points from 2023.

Panel B: What do Americans Think of Government Funding of Science?

Although Americans in both parties largely agree that government investments in science are worthwhile, Democrats and Republicans disagree to a far greater degree if they are essential for scientific progress.

While most Republicans see government investment as worthwhile, they are far less likely to say government investment is essential. A slight majority of Republicans say private investment will ensure enough scientific progress is made, even without government investment (54%).

Democrats tilt to a far greater degree in the other direction, with about eight-in-ten saying government investment in research is essential for scientific progress.

Panel C: Americans' Confidence in Scientists Compared with Other Groups

Scientists have consistently ranked at the top or near the top of the groups since we started asking about confidence in 2016. In the new survey, Americans have similar levels of confidence in scientists and the military. The military has consistently ranked near the top with scientists in our surveys.

Panel C

Americans still have more confidence in scientists than many other groups in society

% of U.S. adults who have ___ of confidence in the following groups to act in the best interests of the public

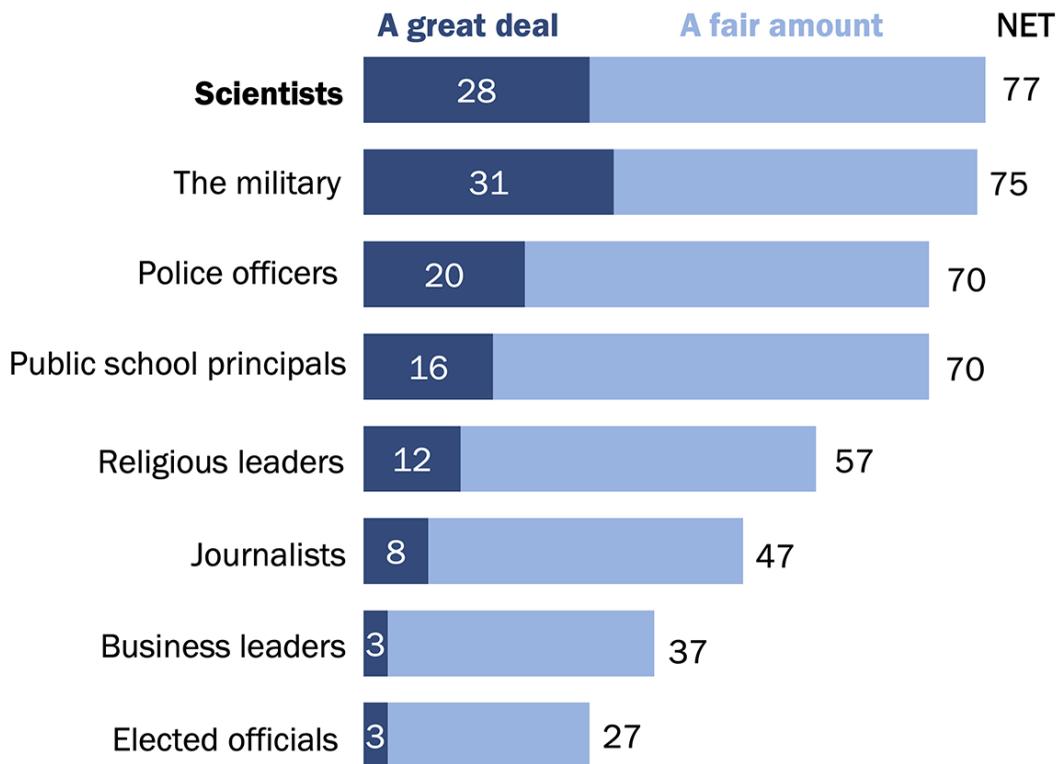


Figure sourced from a survey of U.S. adults conducted Oct. 20–26, 2025. "Do Americans Think the Country is Losing or Gaining Ground in Science?"

Note: Respondents who gave other responses of "Not too much" or "None at all" or did not give an answer are not shown.

Americans have the least amount of confidence in business leaders and elected officials to act in the best interests of the public. These two groups have consistently ranked at the bottom of the groups included in our surveys.

Scientists are not the only group that the public has less confidence in than five years ago. Between 2020 and 2023, confidence in all eight groups we asked about declined. ✨

Pew Research Center

Pew Research Center is made up of intellectually curious people from across many fields, places, and backgrounds. For decades, our expertise, best-in-class methods, and commitment to high-quality, fact-based research have made us a trusted resource for leaders and everyday people around the world. We generate a foundation of facts that enriches public dialogue and supports sound decision-making. We are nonprofit, nonpartisan, and non-advocacy. We value independence, objectivity, accuracy, rigor, humility, transparency, and innovation.

Learn more at www.pewresearch.org.



Medical minds meet here.

We gather thought leaders from around the continent and organize large-scale medical symposia that afford our partners the chance to build relationships with the clinical community and stay up-to-date on the current conversations in multiple disease areas. Ask us how you can be part of our conferences by reaching out to info@catalytichealth.com



Who's Doing What and Who's Going Where

Marie-Josèphe Champagne has returned to **FORUS Therapeutics** Inc as Senior Medical Science Liaison for Quebec and Atlantic Canada.

Carly Epp is starting a new position as VP of Sales and Marketing at **Q&C Services**.

Sabreena Bhanji has taken up a new role as Learning Partner at **AbbVie**.

Paul Zuckernick has started a new position as Cardiovascular Renal Account Specialist at **Bayer**.

Paul McGuckin has begun a new position as Ontario Regional Business Manager at **Pfizer Global Hospital & Biosimilars GH&B**.

Kori Kostka has started a new position as Healthcare Education Manager at **Uresta**.

Darren Breckenridge has embarked on a new position as Director, Competitive Strategy at **Vertex Pharmaceuticals**.

BioScript Solutions is pleased to announce the appointment of **Juan Hoyos** as Vice President, Wholesale & Logistics.

Kerollos Hanna has begun a new position as Medical Advisor, Sleep Medicine at **Eisai Canada**.

Patricia Sarah Olenga has started a new role at **AstraZeneca** as Senior Manager, Digital Customer Strategy.

Rodney Naylor has taken on a new position as Medical Science Liaison at **GSK**.

Jennifer Barbusci is starting a new position as Marketing Manager, Rheumatology at **AbbVie**.

Marc Visconti has joined **BioSyent Inc.** as Senior Marketing Manager.

Andrew Forgione is starting a new position as Executive Director, Global Virology & Inflammation Public Affairs at **Gilead Sciences**.

Angeline Smith has embarked on a new role as Head of Portfolio Operations at **Sanofi**.

Derek Dupuis has started a new position as a Respiratory Sales Representative with **Xediton Pharmaceuticals**.

Fin Whillans has started a new position as Head of Neuroscience Market Access at **Abbvie**.

Simon Reid is embarking on a new role as Territory Manager at **Keir Surgical Ltd.**

Marc Parent is starting a new role as Commercial Director of **BioCryst Pharmaceuticals**

Karen Geoffroy has started a new position as Medical Science Liaison at **Merck Canada**.

Isabelle Dagenais has begun a new position as Business Unit Head, Dermatology, at **Bausch Health**.

Deborah Padden is embarking on a new role as Associate Director, Oncology Marketing at **Kyowa Kirin**.

Patrice Millaire is starting a new role as Director, Business Development at **Imp**.

Adam Louie has taken on a new position as Senior Talent Acquisition Specialist at **Impres**.

Adam Olmstead has embarked on a new position as Cholesterol Business Unit Commercial Lead at **Merck**.

Please submit your selection for our "People on the Move" section, celebrating the advancements of your colleagues, for upcoming issues via email to info@catalytichealth.com.

Hypothesis

PERSPECTIVES INSIGHTS & THOUGHT LEADERSHIP IN THE LIFE SCIENCES

**TO REGISTER FOR AND RECEIVE
FUTURE ISSUES OF
HYPOTHESIS MAGAZINE
PLEASE VISIT
hypothesismag.com**