NEJM AI

NEJM AI 2025;2(6)
DOI: 10.1056/AIoa2300015

ORIGINAL ARTICLE

People Overtrust AI-Generated Medical Advice despite Low Accuracy

Shruthi Shekar, B.S.,[1] Pat Pataranutaporn, Ph.D.,[1] Chethan Sarabu, M.D.,[2,3,4] Guillermo A. Cecchi, Ph.D.,[5] and Pattie Maes, Ph.D.,[1]

Received: June 29, 2023; Revised: July 11, 2024; Accepted: October 21, 2024; Published: May 13, 2025

Abstract

BACKGROUND This article presents a comprehensive analysis of how artificial intelligence (AI)–generated medical responses are perceived and evaluated by nonexperts.

METHODS We conducted a study in which a total of 300 participants gave evaluations for medical responses that were either written by a medical doctor on an online health care platform or generated by a large language model and labeled by physicians as having high accuracy or low accuracy.

RESULTS Results showed that participants could not effectively distinguish between AI-generated responses and doctors' responses and demonstrated a preference for AI-generated responses, rating high-accuracy AI-generated responses as significantly more valid, trustworthy, and complete/satisfactory. Low-accuracy AI-generated responses on average performed very similarly to doctors' responses. Participants not only found these low-accuracy AI-generated responses to be valid, trustworthy, and complete/satisfactory, but also indicated a high tendency to follow the potentially harmful medical advice and incorrectly seek unnecessary medical attention as a result of the response provided. This problematic reaction was comparable with, if not stronger than, the reaction they displayed toward doctors' responses. Both experts and nonexperts exhibited bias, finding AI-generated responses to be more thorough and accurate than doctors' responses but still valuing the involvement of a doctor in the delivery of their medical advice.

CONCLUSIONS The increased trust placed in inaccurate or inappropriate AI-generated medical advice can lead to misdiagnosis and harmful consequences for individuals seeking help. Further, participants were more trusting of high-accuracy AI-generated responses when told they were given by a doctor, and experts rated AI-generated responses significantly higher when the source of the response was unknown. Ultimately, AI systems should be implemented in collaboration with medical professionals when used for the delivery of medical advice in order to prevent misinformation while reaping the benefits of such cutting-edge technology.

Ms. Shekar and Dr. Pataranutaporn contributed equally to this article.

The author affiliations are listed at the end of the article.

Ms. Shekar can be contacted at scshekar@stanford.edu; and Dr. Pataranutaporn can be contacted at patpat@mit.edu.

Introduction

The use of artificial intelligence (AI) in medicine and health care has increased in various domains and applications[1-5] in recent years, from radiology imaging[6] to mental health chatbots[7] and drug discovery.[8] The Covid-19 pandemic further reinforced people's comfort in seeking medical information online, with more accessible means of receiving on-demand medical information.[9-11] Rapid advancement of generative AI has led to large language models (LLMs) and generative pretrained transformers (GPTs), including Gemini, LaMDA, Llama, and Alpaca,[12-18] with capabilities of language generation and question answering in various domains.[16] Researchers have explored the use of LLMs for automating and supporting medical tasks, including diagnosis and triage,[19-21] providing treatment information,[22] treatment prescription,[23] assisting surgery,[24-26] supporting patients' mental health and well-being,[27-33] analyzing laboratory results,[34] medical report generation,[35,36] medical text deidentification,[37] and medical education.[38-41] Recently, Microsoft and Epic Systems have explored the use of LLMs to help decrease administrative burdens on health care providers. University of California San Diego Health; University of Wisconsin Health in Madison, Wisconsin; and Stanford Health Care were among the first organizations to deploy technology to respond to health care messages automatically.[42,43]

Studies have shown notable performances of LLMs completing medical licensing exams.[44-49] One study showed that GPT-4 exceeds the passing score of the official practice materials for the United States Medical Licensing Examination.[44] Another study found that ChatGPT was able to generate higher quality and more empathetic responses to patient questions.[50] A randomized controlled trial for medical diagnosis comparing physicians alone, AI alone, and physicians augmented with AI had the unexpected finding that AI alone outperformed all the other groups.[51] However, a follow-up study from the same group found that physicians augmented with AI performed comparably to AI alone and both groups together outperformed physicians not using AI.[52]

Despite their potential benefit to health care and medicine,[53-55] the stochastic nature of LLMs makes it challenging to determine when LLMs will give factually correct answers versus confidently provide false information (i.e., hallucination or confabulation).[56,57] The stakes are high in medical applications. For instance, a study on the use of LLMs to select next-step antidepressant treatment in major depression showed that the model's inclusion of less optimal clinical recommendations posed a significant risk if used routinely without expert supervision.[23]

As LLMs become more prevalent in mainstream search engines and conversational interfaces, it is not always feasible to have expert supervision. Simply focusing on the accuracy of LLMs in answering medical questions is insufficient, as this fails to capture the broader implications of the technology for the health care system and society at large.[53,58,59] We argue that it is critical to study how the lay public perceives, evaluates, and is affected by AI-generated responses, especially when incorrect, as LLM nonexperts will encounter situations where they might trust AI-generated advice, particularly in the absence of immediate medical professional guidance. Overrelying on false or incomplete AI-generated responses could lead to delayed or inappropriate treatment, potentially worsening health outcomes and even endangering lives.

In this study, we analyzed patient questions from a reputable telemedicine platform to investigate how well participants distinguish between doctor-provided responses and responses generated by AI; how participants rate the validity, trustworthiness, satisfaction, and other aspects of the AI-generated responses compared with doctors' responses; and how participants' knowledge of the response source influences their perception of the response.

Methodology

This article presents three experiments investigating AI-generated medical responses to medical questions and whether the AI-generated responses are comparable to physician responses. Additionally, this study explores the perception of these AI-generated medical responses from the perspective of both the public and physicians.

DATASET GENERATION

One hundred and fifty anonymous medical questions and doctors' responses were retrieved from the question forum on HealthTap (https://www.healthtap.com/), an online health care provider. The inquiries cover six domains of medicine: preventative and risk factors; conditions and symptoms; diagnostics and tests; procedures and surgeries; medication and treatments; and recovery and wellness, with equal distribution.

Using GPT-3,[12] AI responses were produced for each of the medical questions. We applied the default settings of

GPT-3 without any modification to assess the performance of the baseline model (model: text-davinci-002, temperature: 0.7, maximum length: 256).

These AI-generated responses were subsequently evaluated by four evaluators to establish the ground truth on whether the AI-generated responses were accurate. The evaluators were general physicians recruited from Stanford University and University of California, San Francisco, who were compensated with a $100 Amazon gift card for their contribution. Each expert evaluator was asked to evaluate the responses on three axes: accuracy, completeness, and strength. While we primarily focused on the accuracy ratings based on initial clinical testing, it was important to provide these additional dimensions of evaluation to ensure we had the best possible accuracy ratings. For example, an AI-generated response to an inquiry about treatment for a condition may provide an accurate response but exclude alternative treatments. When asked to evaluate an answer, if the axis of completeness was not available the reviewer may have a more difficult time saying this answer was accurate, as it omitted other details; however, by providing these axes of evaluation the reviewer could say that this answer is accurate but incomplete.

Experts were informed that they would be evaluating AI-generated responses, thus making this a nonblind evaluation. Based on these evaluations made by four practicing physicians, providing the ground truth on whether the AI-generated response was correct (yes = 3, maybe = 2, no = 1), compiled scores for each medical response in a dataset of a 150 medical question–response pairs were formed. The medical responses were then organized according to the different accuracy levels. Any response with two or less "maybe" evaluations was considered to have high accuracy. Any inquiry with the majority of evaluations being rated maybe or worse was considered to have low accuracy. This organized dataset was utilized to conduct an array of studies. We used the experts' ratings to select a subset of medical question–response pairs, forming a new dataset consisting of 30 distinct pairs of medical questions and doctors' responses; 30 distinct pairs of medical questions and high-accuracy AI-generated responses; and 30 distinct pairs of medical questions and low-accuracy AI-generated responses.

Later, six additional physicians were asked to evaluate the same medical question–response pairs but with the source of the response unknown (blind evaluation). Given evidence of the response source influencing participant evaluation of AI-generated responses, we were interested in exploring if our physicians also revealed any particular biases during blind evaluations of AI-generated responses. Therefore, three of the experts were asked to evaluate dataset A, a dataset of 150 medical question–response pairs with a random selection of 75 inquiries answered by a doctor and the other 75 by AI. The remaining three experts received dataset B, where the same random selection of 75 inquiries were this time answered by AI and the remaining inquiries were answered by a doctor.

There was a total of three expert evaluations for each of the 150 AI-generated question–response pairs and 150 doctors' response pairs in both the blind and nonblind experiments. This amounted to a sample of 150 composite expert evaluation scores for AI-generated responses and 150 for doctors' responses in each of the evaluations (blind vs. nonblind).

TASK DESCRIPTION

First, we investigated whether participants would be able to distinguish AI-generated responses from doctors' responses as a preliminary understanding of participant perception of AI and doctors in responding to health inquiries. In this first experiment, 100 online participants were presented with 10 medical question–response pairs randomly selected from a collection of 30 doctors' responses, 30 high-accuracy AI-generated responses, and 30 low-accuracy AI-generated responses. After reading the provided medical question–response pair, participants provided Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) on their understanding of the medical question and their understanding of the response. Additionally, they indicated their belief about the response source (response given by a doctor or an AI text generator) and provided a Likert scale evaluation of their confidence in the source they selected on a scale of 1 (low confidence) to 5 (high confidence). The full set of questionnaires is listed in the Supplementary Appendix.

For the second experiment, we assessed how participants evaluate responses generated by the AI system compared with those provided by doctors, when they are unaware of the exact source of the responses. This experiment, similarly to experiment 1, involved 100 participants, who were presented with 10 medical question–response pairs randomly selected from a collection of doctors' responses, high-accuracy AI-generated responses, and low-accuracy AI-generated responses. Here, participants provided Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) on their understanding of the medical question and their understanding of the response. Additionally, they were asked to indicate their perception of response validity (yes/no). Finally, participants provided Likert

scale evaluations of the trustworthiness of the response; the completeness and satisfaction of the response; participant tendency to search for additional information based on the response; participant tendency to follow the advice provided in the response; and participant tendency to seek subsequent medical attention as a result of the response.

In the third experiment, we investigated if participants exhibited biases toward or against certain response types. Similarly to experiments 1 and 2, 100 participants were presented with 10 medical question–response pairs randomly selected from a collection of doctors' responses, high-accuracy AI-generated responses, and low-accuracy AI-generated responses. However, at the start of the survey, participants were randomly shown one of three labels: "The responses to each medical question were given by a %(doctor)"; "The responses to each medical question were given by %(artificial intelligence (AI))"; or "The responses to each medical question were given by a %(doctor assisted by AI)." Participants then provided Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) on their understanding of the medical question and their understanding of the response. They indicated their perception of response validity (yes/no). Finally, they provided Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) of the trustworthiness of the response; the completeness and satisfaction of the response; participant tendency to search for additional information based on the response; participant tendency to follow the advice provided in the response; and participant tendency to seek subsequent medical attention as a result of the response.

LINGUISTIC ANALYSIS

Additionally, linguistic analysis was completed to identify if there were any significant variations in linguistic characteristics (i.e., word count, sentiment, and readability) in the different response types. We used the Valence Aware Dictionary and Sentiment Reasoner (VADER) sentiment analysis tool attuned to sentiments expressed in social media.[60] We used VADER sentiment scores based on the vaderSentiment library and Flesch Reading Ease readability score based on readability library.[61,62] We used descriptive statistics and the analysis of variance (ANOVA) to analyze the results.

PARTICIPANTS

We recruited the participants using the website Prolific. Participants were prescreened to be fluent in English and older than 18 years of age. The study was set to be balanced between male and female participants.

We excluded participants who did not complete the full study and did not pass the screening questions and attention check. After exclusions, we had 98 participants for experiment 1, 96 participants for experiment 2, and 100 participants for experiment 3. The full demographic data are in the Supplementary Appendix.

STATISTICAL ANALYSIS

The 30 medical responses from the three conditions (doctor, high-accuracy AI, and low-accuracy AI) each had approximately 8 to 12 participant evaluations. As a result, each condition had a total of approximately 300 participant evaluations. For experiments 1, 2, and 3, linear mixed effects models, with crossed random effects for subjects and question–response pairs, were used to account for the fact that the participants rated multiple question–response pairs from different response types in a random order. This approach allowed for the analysis of multiple evaluations of different medical responses completed by the same participant. The models were constructed in R (version 4.3.1) using the lme4 package, with the response scores as the dependent variable. An omnibus test of whether average response scores were the same across the three conditions was conducted using a likelihood ratio test. Pairwise comparisons among conditions were calculated using the emmeans package in R, with P values adjusted for family-wise error using the sequential Bonferroni method. The intraclass correlation was calculated using the psych package in R, with the adjusted intraclass correlation reported. For the additional experiment on the physicians' evaluations of AI-generated responses with and without response source shown, a simple t-test was utilized to analyze significance between response sources (AI-generated vs. doctor) within each of the groups (accuracy ratings from the blind study, accuracy ratings from the nonblind study, and so on). Additionally, to identify if there were statistically significant differences between scores arising across the different test types (blind vs. nonblind) and different response sources (AI-generated vs. doctor), a two-way ANOVA test was completed.

Results

One hundred and fifty anonymous medical questions and their corresponding doctors' responses were retrieved from HealthTap. AI-generated responses were produced for each inquiry and evaluated by four practicing physicians from Stanford Health and University of California, San Francisco, who rated accuracy as yes, maybe, or no.

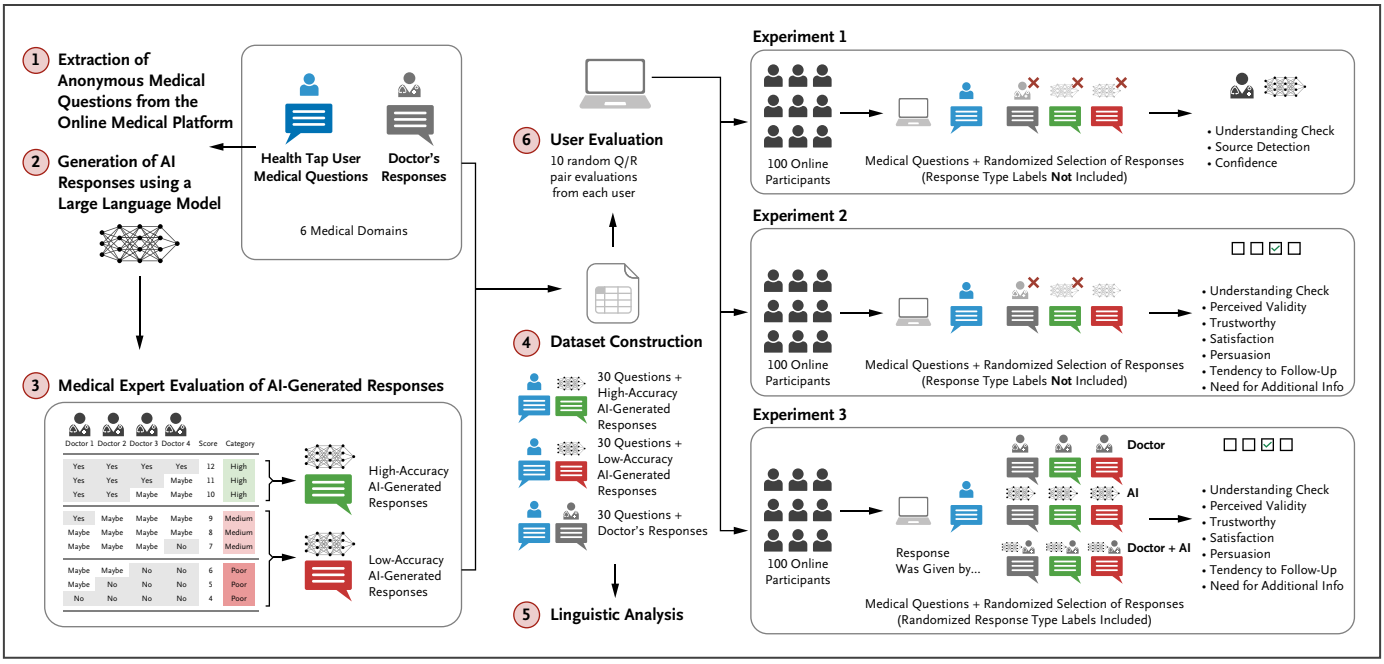

Figure 1. Visual Summary of the Dataset Construction and Pipeline of Experiments.

AI denotes artificial intelligence; info, information; Q, question; and R, response.

Responses with two or fewer maybe evaluations and zero no evaluations were considered high accuracy, while those with a majority of maybe or worse responses were considered low accuracy; see Figure 1 and Figure 2. Of the 150 AI-generated responses, 56.0% achieved high accuracy and 44.0% low accuracy; see Figure 3. Using these expert ratings, we created a dataset of 90 question–response pairs: 30 with doctors' responses, 30 with high-accuracy AI responses, and 30 with low-accuracy AI responses. We then conducted studies with 100 online participants recruited through a platform designed to represent the general public (demographic data are in the Supplementary Appendix) to assess the perception of AI-generated responses versus doctors' responses. Detailed methodology and analysis are provided in the Supplementary Appendix.

EXPERIMENT 1: EVALUATION OF PARTICIPANT ABILITY TO DISTINGUISH ARTIFICIAL INTELLIGENCE–GENERATED RESPONSES FROM DOCTORS' RESPONSES

First, we investigated whether participants would be able to distinguish AI-generated responses from doctors' responses as a preliminary understanding of participant perception of AI and doctors' performance in responding to health inquiries. To do so, participants were provided with a medical question and a corresponding response, a doctor's response, a high-accuracy AI-generated response, or a low-accuracy AI-generated response. To reiterate, as judged by our four expert evaluators, high-accuracy AI refers to responses that are generated by the AI system with a high degree of accuracy, while low-accuracy AI refers to responses generated by the AI system with a lower degree of accuracy; see Figure 3.

Overall, in experiment 1, 100 online participants (98 participants passed the screening and were included in the result) were presented with 10 randomly selected medical question–response pairs from a collection of 30 doctors' responses, 30 high-accuracy AI-generated responses, and 30 low-accuracy AI-generated responses. After reading the provided medical question–response pair, participants were asked to provide Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) of their understanding of the medical question and their understanding of the response. Additionally, they were asked to indicate their belief about the response source (response given by a doctor or an AI text generator) and provide a Likert scale

Medical Domain	Question	Doctor's Response	AI-Generated Response
Preventative and Risk Factors	I'm an offensive lineman that wants to prevent hand injury by taping my hands and fingers. How should I tape them to get the most benefit?	Tape is usually : Applied one of two ways. (first make sure its not too tight and its better to have a trainer do it if available and the tape should not roll or be too tight or be too sticky) either single tape to a joint on one digit or buddy tape to keep an injured finger next to another. If you buddy tape often you avoid the joints and pad the area lightly between the fingers, if you single tape a figure 8. Make sure the tape doesnt roll up on itself, if it does you can gt a tight ring that tightens as you move and you'll end up needing to cut it off when circulaiton is compromised. Tape is no subsititute for a swollen injured digit, it there is an injury seek medical attention.	There are a few ways that you can tape your hands and fingers to prevent hand injury. One way is to tape your fingers together so that they are unable to move independently. This will help to prevent them from being bent the wrong way or getting caught on something. Another way is to tape your palms so that they are protected from getting hit. This will help to prevent bruising and other hand injuries.
Conditions and Symptoms	My 3 month old has a cough. When should I worry? He's exclusively breasfed, is eating well and has no fever. Sometimes has nasal congestion but we suck it out with Saline and Nose Frida Device	Basics: If the kid feeds vigorously & retains it's normal activity level things are likely fine. If the feeding falls off, the kid seems to struggle to feed or breathe after nasal suction things are moving in the direction where an evaluation soon is needed. Sometimes this is just reaction to airborne irritants but it could be RSV or other worrisome problem.	If your baby is coughing and otherwise acting normally, there is no need to worry. However, if the cough is accompanied by other symptoms such as fever, difficulty breathing, or lethargy, you should seek medical attention.
Diagnostic and Tests	Had a traumatic brain injury 4 yrs ago. been trying to cure my constant headache. I think I may have occipital neuralgia. How would I diagnose this? Thanks	PCP...Neurologist: 17 y/o female says, Had a traumatic brain injury 4 yrs ago. Been trying to cure my constant headache. I think I may have occipital neuralgia. How would I diagnose this? Thanks. Injury seemed to have happened when you were technically pediatric/ Adolescence age group. Should first see the PCP or Pediatric Neurologist for evaluation before anything else. The sooner the better. Wish you well soon.	There is no one definitive test for diagnosing occipital neuralgia. A diagnosis is typically made based on a combination of a thorough medical history, a physical examination, and additional tests such as imaging studies (e.g., MRI, CT scan) and/or nerve conduction studies/electromyography.
Procedures and Surgeries	Will getting lung reduction surgery fix my subcutaneous emphysema?	No: Subcutaneous emphysema is usually the result of a collapsed lung (pneumothorax). It is treated with placing a drainage tube between the ribs to "suck" the air out from around the collapsed lung and allow it to reexpand. Subcut. Emphysema can take days to resolve even after the lung expands. Lung volume reduction surgery is really an elective procedure rather than treatment for an acute process.	It is possible that lung reduction surgery may improve subcutaneous emphysema, but this is not guaranteed.
Medications and Treatments	Does medical marijuana, with high cannabidiol and low THC content, have negative interactions with lisinopril, hydrochlorothiazide, and a pacemaker that performs 87% of an elderly person's heart beat?	2 difficulties with: this question. Pacemakers do not "perform" heart beats. They only emit tiny electrical shocks to heart muscle when heart contraction not electrically detected within time limit set in pacemaker programming. Electronic pacemaker initiated heart beats are always weaker because less well coordinated. No clinical research trial data exists to answer 1st questions; though zero basis to believe healthy.	There is no definitive answer to this question since there is limited research on the potential interactions between medical marijuana and lisinopril, hydrochlorothiazide, and pacemakers. However, it is generally recommended that people consult with a healthcare provider before using medical marijuana, especially if they are taking other medications. Additionally, it is important to be aware that marijuana can interact with other medications and medical devices, so it is important to use caution and consult with a healthcare provider if there are any concerns.

Figure 2. Example Medical Questions by Category: Comparing Doctor-Written and Artificial Intelligence–Generated Responses.

AI denotes artificial intelligence; CT, computed tomography; MRI, magnetic resonance imaging; PCP, primary care provider; RSV, respiratory syncytial virus; THC, tetrahydrocannabinol; and y/o, years old.

evaluation of their confidence in the source they selected on a scale of 1 (low confidence) to 5 (high confidence); see Figure 4. The full set of questionnaires is listed in the Supplementary Appendix.

Understanding Inquiry and Response

From the hierarchical linear model analysis, there were no significant differences in the participants' understanding of medical questions (P=0.5964) across the three categories (Fig. 5). However, significant differences were observed in participant evaluations of response understanding ($P=1.787\times10^{-4}$). From the pairwise comparison, we found that participants rated their understanding of AI-generated responses, regardless of the accuracy level (high-accuracy AI: mean = 4.53, standard deviation [SD] = 0.69; low-accuracy AI: mean = 4.56, SD = 0.62), to be significantly higher (high-accuracy AI vs. doctors: P=0.001; low-accuracy AI vs. doctors: P=0.0008) than their understanding of doctors' responses (mean = 4.32, SD = 0.83).

Source Determination Accuracy

When participants were asked to determine the source of the medical response provided to them (doctor-written or AI-generated), there was an average source determination accuracy of 50% for doctors' responses, 53% for high-accuracy AI responses, and 50% for low-accuracy AI responses. There were no significant differences in the participants' performance in the source determination task (P=0.6539)

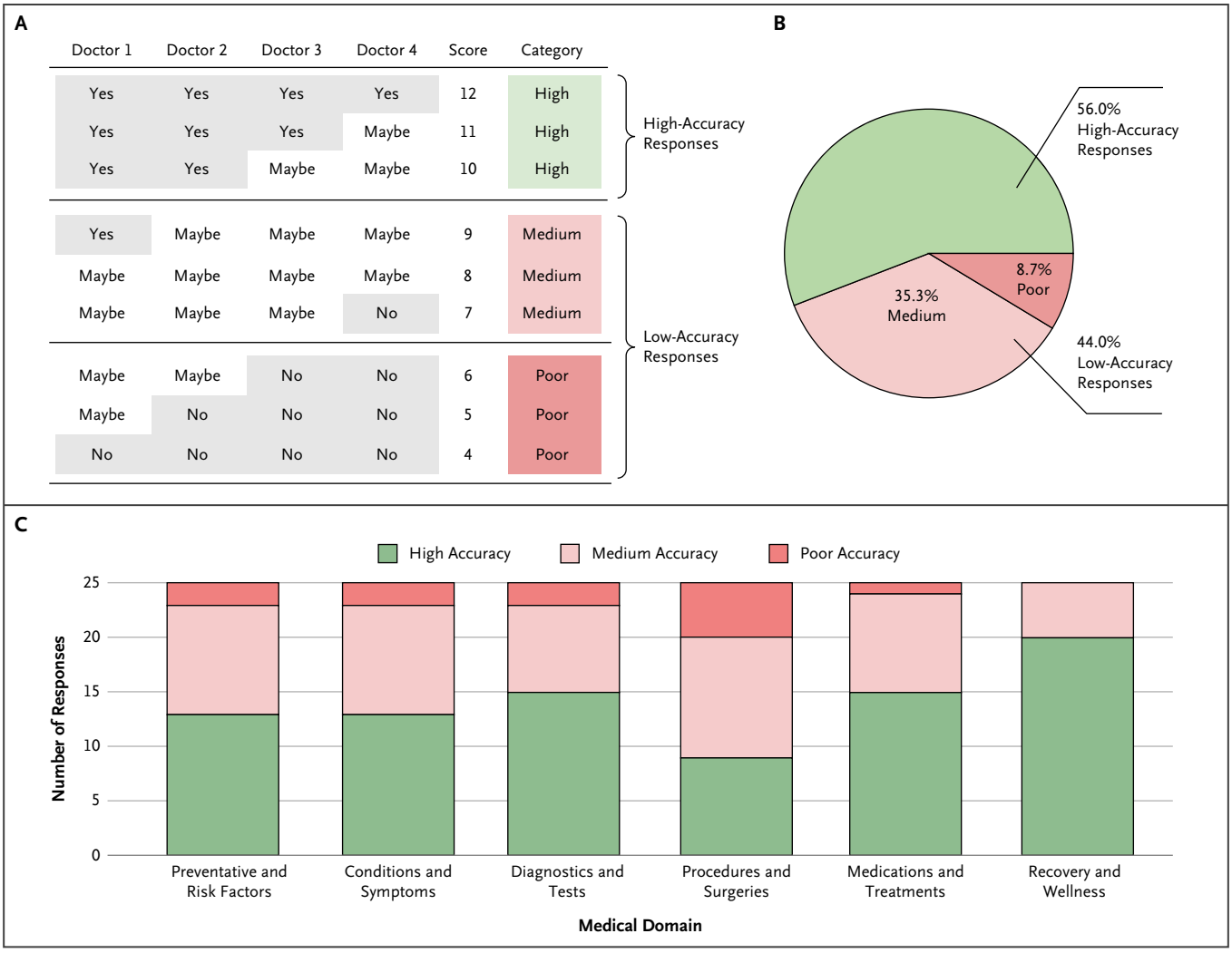

Figure 3. Expert Evaluation of Artificial Intelligence–Generated Medical Response Accuracy.

In Panel A, the table represents the compilation of the four physicians' accuracy evaluation scores, with the values for each evaluation as follows: yes = 3, maybe = 2, no = 1. Using the following numerical values for each expert evaluation, a compiled score was formed. Any score equal to or above 10 (with two or fewer maybe evaluations) was considered high accuracy. Any score equal to or below 9 (majority of evaluations are maybe or worse) was considered low accuracy. In Panel B, in a dataset of 150 artificial intelligence–generated medical responses, 56.0% were of high accuracy and 44.0% were of low accuracy. Panel C shows a breakdown of high- and low-accuracy artificial intelligence–generated responses across the six different medical domains.

across the three categories, indicating that participants were unable to effectively distinguish doctors' responses from AI-generated medical responses.

Source Determination Confidence

When asked to provide their level of confidence in determining the source of the medical response provided to them, even though their accuracy was low (around 50%), participants reported a high level of confidence across the three types of response when they answered correctly (doctors' response: mean = 3.94, SD = 0.87; high-accuracy AI response: mean = 3.78, SD = 0.91; low-accuracy AI response: mean = 3.94, SD = 0.79) and incorrectly (doctors' response: mean = 3.98, SD = 0.84; high-accuracy AI response: mean = 3.77, SD = 0.93; low-accuracy AI response: mean = 4.02, SD = 0.84). The level of confidence when participants guessed correctly and incorrectly was not significantly different across the three response types (doctors' response: P=0.6803; high-accuracy AI response: P=0.9279; low-accuracy AI response: P=0.9537).

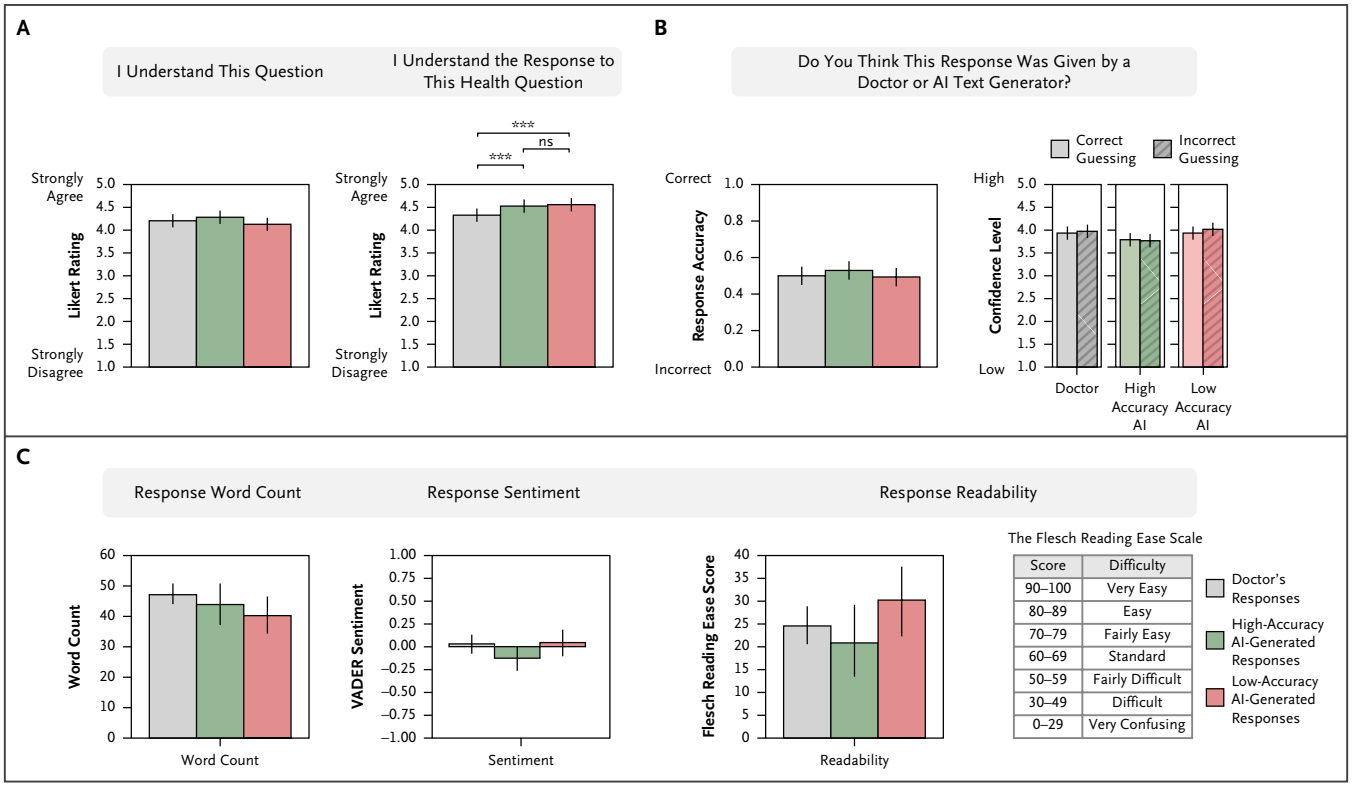

Figure 4. Participants' Ability to Distinguish between Artificial Intelligence–Generated and Doctor-Written Medical Responses.

Panel A shows the Likert analysis of participant understandability of each medical question and response. Panel B shows the quantification of responses perceived as artificial intelligence (AI)–generated versus provided by a doctor in the different response types (doctor, high-accuracy AI, low-accuracy AI) with participant confidence levels. Panel C shows analysis of medical response word count, sentiment, and readability. AI denotes artificial intelligence and ns, not significant.

Linguistic Analysis

We found that there were no significant differences identified in word count (P=0.154; doctors' responses: mean = 46.9, SD = 20.3; high-accuracy AI response: mean = 43.6, SD = 30.1; low-accuracy AI response: mean = 40.0, SD = 26.4), VADER sentiment[60] value (P=0.107; doctors' responses: mean = 0.0312, SD = 0.626; high-accuracy AI response: mean = −0.128, SD = 0.590; low-accuracy AI response: mean = 0.0460, SD = 0.559), and Flesch Reading Ease readability score[61,62] (P=0.250; doctors' responses: mean = 24.6, SD = 26.8; high-accuracy AI response: mean = 20.8, SD = 36.7; low-accuracy AI response: mean = 30.4, SD = 33.2) in the different response types.

EXPERIMENT 2: PARTICIPANTS' EVALUATION OF ARTIFICIAL INTELLIGENCE–GENERATED RESPONSES COMPARED WITH DOCTORS' RESPONSES

Experiment 2 aimed to assess how participants evaluate responses when unaware of the source. One hundred participants (96 participants passed the screening and were included in the results) were presented with 10 medical question–response pairs randomly selected from a collection of 30 doctors' responses, 30 high-accuracy AI-generated responses, and 30 low-accuracy AI-generated responses.

Here, participants were asked to provide Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) of their understanding of the medical question and their understanding of the response. Additionally, they were asked to indicate their perception of response validity (yes/no). Finally, participants were asked to provide Likert scale evaluations of the trustworthiness of the response; the completeness and satisfaction of the response; participant tendency to search for additional information based on the response; participant tendency to follow the advice provided in the response; and participant tendency to seek subsequent medical attention as a result of the response.

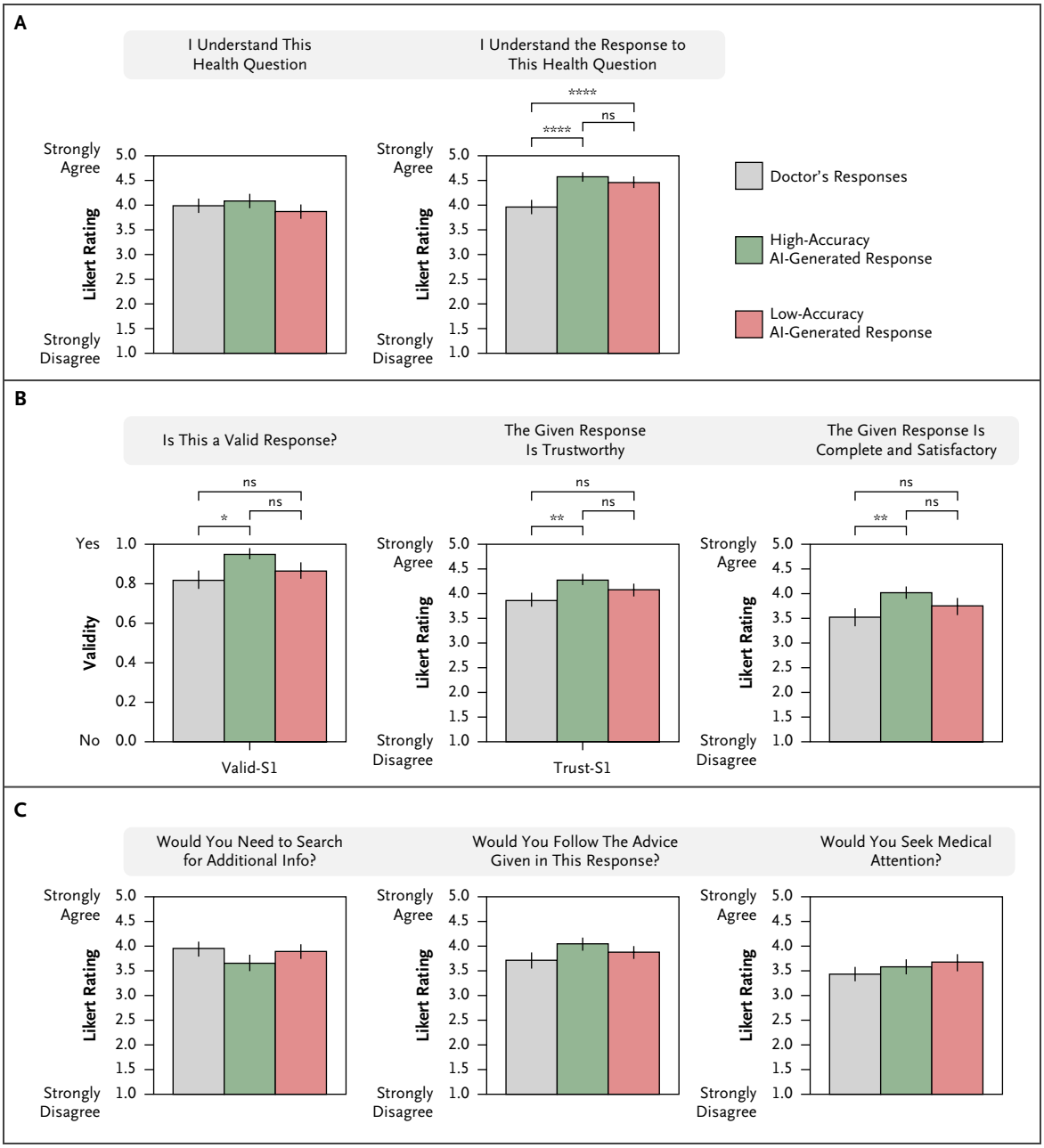

Figure 5. Participants' Perception and Evaluation of Doctor-Written, High-Accuracy Artificial Intelligence–Generated, and Low-Accuracy Artificial Intelligence–Generated Medical Responses.

Panel A shows analysis of participant understandability for each medical question and response. Panel B shows participant evaluation of perceived response validity, trustworthiness, and completeness/satisfaction. Panel C shows analysis of participant responses to survey questions inquiring about tendency to follow-up, including the likelihood of requiring additional information, following the advice provided in the medical response, and seeking subsequent medical attention. AI denotes artificial intelligence and ns, not significant.

Understanding Inquiry and Response

From the hierarchical linear model analysis, there were no significant differences in the participants' understanding of medical questions across the three categories: doctor-written, high-accuracy AI-generated, and low-accuracy AI-generated ($P=0.43$) responses. However, participants demonstrated a significantly higher level of understanding ($P=8.2\times10^{-6}$) of AI-generated responses

than the doctors' responses, regardless of the AI-generated response accuracy level. Participants indicated the highest level of understanding for high-accuracy AI-generated responses (mean = 4.58, SD = 0.73), followed by low-accuracy AI-generated responses (mean = 4.48, SD = 0.87), and then the doctors' responses (mean = 3.97, SD = 1.21) (high-accuracy AI vs. doctor: P<0.0001; low-accuracy AI vs. doctor: P<0.0001).

Validity

Additionally, significant differences were observed in participant evaluations of response validity within the different response types (P=0.011). The pairwise analysis indicated that participants perceived the high-accuracy AI-generated (mean = 0.95, SD = 0.22) responses to be significantly more valid (P=0.0106) than the doctors' responses (mean = 0.81, SD = 0.39). The low-accuracy AI-generated responses (mean = 0.87, SD = 0.34) performed very comparably with the doctors' responses.

Trustworthiness

Significant differences were observed in participant evaluations of response trustworthiness within the different response types (P=0.0058). The pairwise analysis indicated that participants perceived the high-accuracy AI responses (mean = 4.26, SD = 0.86) to be significantly more trustworthy (P=0.0050) than the doctors' responses (mean = 3.85, SD = 1.13). The low-accuracy AI-generated responses (mean = 4.06; SD = 1.05) were rated similarly to the doctors' responses.

Completeness/Satisfaction

Significant differences were observed in participant evaluations of response completeness and satisfaction in the different response types (P=0.005). The pairwise analysis indicated that participants perceived the high-accuracy AI responses (mean = 4.03, SD = 1.11) to be significantly more complete/satisfactory (P=0.0042) than the doctors' responses (mean = 3.55, SD = 1.33). The low-accuracy AI-generated responses (mean = 3.77, SD = 1.30) were rated similarly to the doctors' responses, with no significant difference identified.

Tendency to Seek Additional Information

Beyond the previous metrics, such as validity, trustworthiness, and completeness, we were also interested in gaining an understanding of what next steps the participant might be inclined to take as a result of the response. Participants were told to picture themselves asking their doctor the given question and were asked to rate their tendency to seek additional information as a result of the response they received. We did not observe significant differences (P=0.10) between doctors' responses (mean = 3.94, SD = 1.17), high-accuracy AI responses (mean = 3.65, SD = 1.35), and low-accuracy AI responses (mean = 3.88, SD = 1.21).

Tendency to Follow the Advice Provided

Asked to envision themselves as the patient seeking advice, participants rated their tendency to follow the advice provided to them in the response. We did not observe significant differences (P=0.094) between doctors' responses (mean = 3.68, SD = 1.20), high-accuracy AI responses (mean = 4.00, SD = 1.02), and low-accuracy AI responses (mean = 3.85, SD = 1.14), demonstrating a relatively equal tendency to follow the advice provided across all three response types.

Tendency to Seek Further Medical Attention

Finally, participants were asked to rate their tendency to seek subsequent medical attention as a result of the response provided. We did not observe significant differences (P=0.26) between doctors' responses (mean = 3.42, SD = 1.31), high-accuracy AI responses (mean = 3.56, SD = 1.24), and low-accuracy AI responses (mean = 3.66, SD = 1.28).

EXPERIMENT 3: PARTICIPANTS' EVALUATION OF ARTIFICIAL INTELLIGENCE–GENERATED RESPONSES COMPARED WITH DOCTORS' RESPONSES GIVEN A RANDOM LABEL

We investigated if participants exhibited biases toward or against certain response sources. Similarly to Experiment 2,100 participants (all 100 participants passed the screening and were included in the result) were presented with 10 medical question–response pairs randomly selected from a collection of 30 doctors' responses, 30 high-accuracy AI-generated responses, and 30 low-accuracy AI-generated responses. However, at the start of the survey, participants were randomly shown one of three labels: "The responses to each medical question were given by a %(doctor)"; "The responses to each medical question were given by %(artificial intelligence (AI))"; or "The responses to each medical question were given by a %(doctor assisted by AI)."

Then, similarly to experiment 2, participants were asked to provide Likert scale evaluations on a scale of 1 (strongly

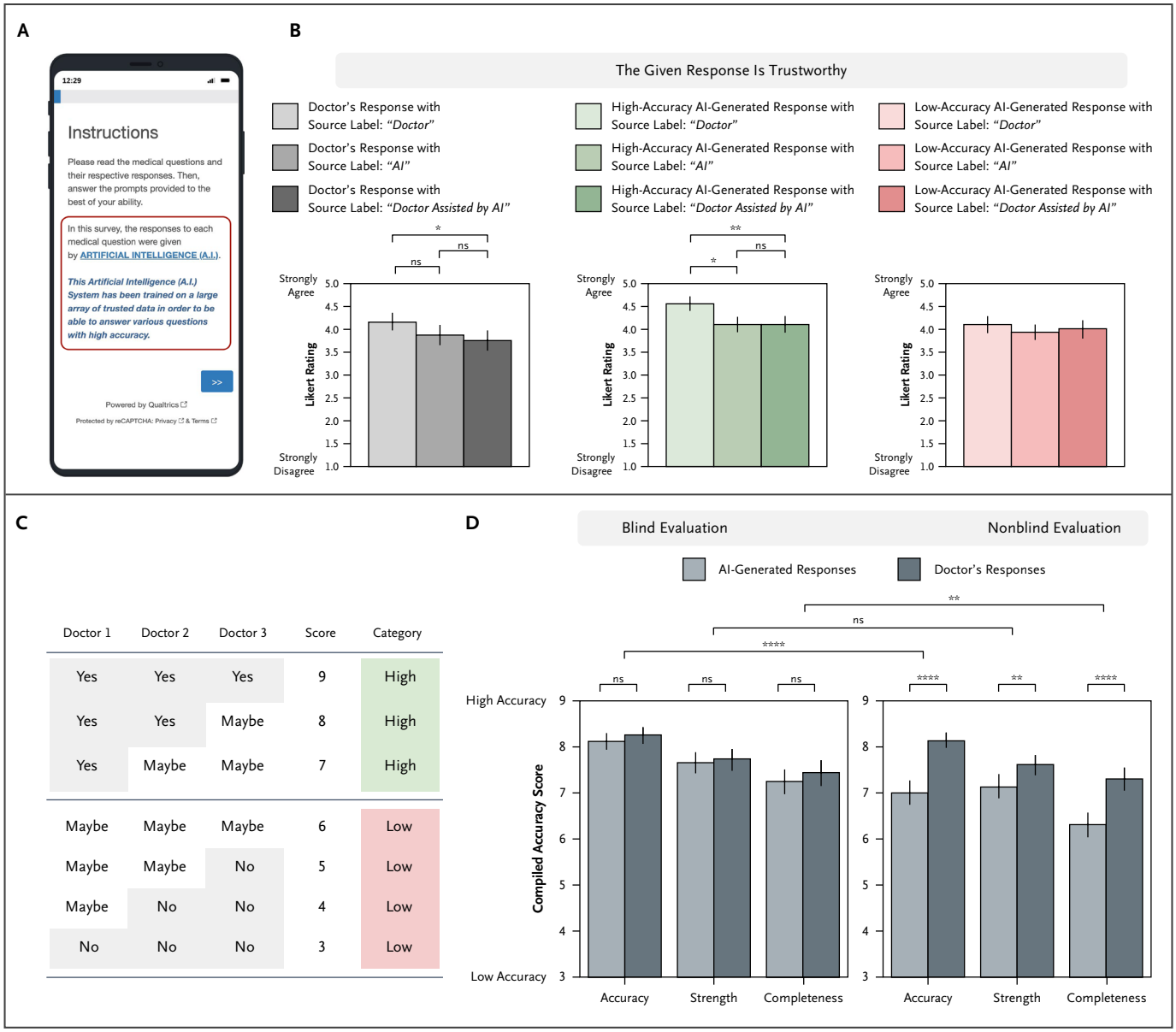

Figure 6. Participant Trust and Expert Evaluation of AI-Generated Medical Responses by Accuracy and Source Label.

Panel A shows a screenshot of the survey platform displaying the randomly assigned source label. Panel B shows Likert analyses of participant evaluations of response trustworthiness across the different response types and different source labels (doctor, AI, and doctor assisted by AI), demonstrating the impact of randomized source labels on participants' perception of doctor-written, high-accuracy artificial intelligence (AI)–generated, and low-accuracy AI-generated medical responses. Panels C and D are part of the blind and nonblind evaluations of the AI-generated medical responses by physicians. In Panel C, the table represents the compilation of three expert evaluation scores with the values for each evaluation as follows: yes = 3, maybe = 2, no = 1. Using these numerical values for each expert evaluation, a compiled score was formed. Any score equal to or above 7 (with two or fewer maybe evaluations) is considered high accuracy. Any score equal to or below 6 (the majority of evaluations are maybe or worse) is considered low accuracy. There was a total of three expert evaluations in each evaluation round, blind and nonblind. Panel D shows compiled accuracy scores for blind evaluation – average of 150 compiled scores from each test (accuracy, strength, and completeness) across both response types (AI generated and doctor written) – and nonblind evaluation – average of 150 compiled scores from each test (accuracy, strength, and completeness) across both response types (AI generated and doctor written). AI denotes artificial intelligence and ns, not significant.

disagree) to 5 (strongly agree) of their understanding of the medical question and their understanding of the response. They were asked to indicate their perception of response validity (yes/no). Finally, they were asked to provide Likert scale evaluations on a scale of 1 (strongly disagree) to 5 (strongly agree) of the trustworthiness of the response; the completeness and satisfaction of the response; participant tendency to search for additional information based on the response; participant tendency to follow the advice provided in the response; and participant tendency to seek subsequent medical attention as a result of the response.

In general, the source labels had little effect on participants' evaluations of the medical responses. However, we observed the effect of the labels on the trustworthiness rating of the doctors' responses (P=0.022) and high-accuracy AI responses (P=0.0042); see Figure 6. In particular, the pairwise analysis revealed that in the presence of the label "This response to each medical question was given by a %(doctor)," participants tended to rate high-accuracy AI-generated responses as significantly more trustworthy (doctor vs. AI: P=0.013; doctor vs. doctor assisted by AI: P=0.01). However, we did not see such an impact of the same label (doctor) on the trustworthiness ratings of the low-accuracy AI-generated response (P=0.49).

ADDITIONAL EXPERIMENT: PHYSICIANS' EVALUATIONS OF ARTIFICIAL INTELLIGENCE–GENERATED RESPONSES WITH AND WITHOUT RESPONSE SOURCE INDICATED

Identifying key results across nonexpert participant evaluations of the AI-generated responses versus the doctors' responses, we wanted to conduct a preliminary investigation of whether similar trends would be found among the physician evaluators. Particularly, we were interested in exploring if our physicians revealed any particular biases during their evaluation of the AI-generated responses. To do so, we first asked three of our four physicians from the initial evaluation to also evaluate the doctors' responses, completing the nonblind portion of the study. Simultaneously, six additional general physicians from the same institutions were asked to complete a blind evaluation of the same AI-generated responses and doctors' responses. More details regarding the design of this additional study can be found in the methodology. We found that when the experts did not have access to the label regarding the source of the response (doctor written or AI generated), there was no significant difference in their evaluation in terms of accuracy (P=0.2258), strength (P=0.5694), and completeness (P=0.2740). However, when the experts did have access to the source of the response, they evaluated the AI-generated responses as significantly lower in all three metrics: accuracy ($P=6.509\times10^{-13}$), strength (P=0.003), and completeness ($P=1.606\times10^{-8}$). Additionally, when completing a two-way ANOVA test, a significant relationship between the study type (blind vs. nonblind) and the response source (AI vs. doctor) was identified while evaluating the accuracy ($P=1.385\times10^{-7}$) and completeness (P=0.001126), confirming a bias presented by experts against AI-generated responses when the source of the response is indicated; see Figure 6.

Discussion

Participants demonstrated similar understanding of medical questions across all groups (high-accuracy AI, low-accuracy AI, and doctor). This consistency, combined with no differences in linguistic characteristics between response types, thus controlling for any confounding factors related to medical question–response linguistics that could impact evaluation outcomes, ensures that evaluation differences reflect perception of responses rather than question comprehension.

PUBLIC TRUST, RISK, AND PERCEPTION OF ARTIFICIAL INTELLIGENCE–GENERATED MEDICAL RESPONSES

Participants displayed an approximate 50% accuracy rate in discerning the origin of the medical responses, making it clear that they struggled to effectively differentiate between medical advice offered by a doctor and medical responses generated by AI. This holds true even when the accuracy of the AI-generated medical response is comparatively low. Thus, participants perceived the AI-generated responses as remarkably similar to those provided by doctors, rendering them unable to accurately differentiate between the advice given by the AI and that offered by a registered physician on the online health care platform HealthTap.

In addition to participants' inability to distinguish AI-generated responses from doctors' responses, we found that participants evaluated AI-generated responses as almost equal to, if not better than, responses provided by doctors across all metrics. AI-generated medical responses were found to be as comprehensive, valid, trustworthy, complete/satisfactory, and persuasive as doctors' responses, with AI-generated responses of high accuracy performing significantly better in a majority of the metrics. Furthermore, on average, albeit not significantly,

low-accuracy AI-generated responses presented a higher level of performance than the doctors' responses across all the evaluation metrics.

Participants' inability to differentiate between the quality of AI-generated responses and doctors' responses, regardless of accuracy, combined with their high evaluation of low-accuracy AI responses, which were deemed comparable with, if not superior to, doctors' responses, presents a concerning threat. When unaware of the response's source, participants are willing to trust, be satisfied, and even act upon advice provided in AI-generated responses, similarly to how they would respond to advice given by a doctor, even when the AI-generated response includes inaccurate information. This unexpected trust and satisfaction with low-accuracy AI-generated responses may lead to unwitting acceptance of harmful or ineffective medical advice and concerns of liability for any resulting adverse patient outcomes.[63]

Participants evaluating unlabeled medical responses favored AI-generated ones, trusting even low-accuracy responses. However, source labeling changed evaluations significantly. High-accuracy AI responses labeled as doctor were deemed more trustworthy than when labeled as AI, suggesting that, while participants appreciate AI-generated advice, they generally prefer receiving it from doctors. Notably, the doctor label alone did not enhance perception of low-accuracy AI responses. This effect was strongest with high-accuracy AI responses, demonstrating a combined effect of desirable source and high-accuracy model for achieving desirable evaluations. Similar patterns appear in other domains, as shown in a study where humanlike explanations combined with high-accuracy AI responses increased trust in a legal decision-making advice.[64] Interestingly, our expert evaluators showed similar bias, rating AI responses significantly higher when source-blind. This reveals that even those responsible for establishing objective truth and assessing model efficacy can be susceptible to inherent biases.

EXTENDING THE APPLICABILITY OF OUR FINDINGS TO OTHER LANGUAGE MODELS

We used AI-generated responses from the GPT-3 model, which is among the most adopted language models with publicly accessible specifications and training data.[12] While there are recent models with greater accuracy, including closed-source models, such as GPTs Claude and Gemini, and open-source alternatives like Llama, we believe that our findings regarding nonexperts' perception and evaluation of AI-generated medical responses can be generalized to AI-generated responses from other advanced language models. This is due to the shared underlying architectures and training methodologies across these models that permit newer models to hallucinate misinformation still.[1,6,44,53,57] Regardless of the specific language model employed, the possibility of generating both highly accurate and inaccurate medical responses remains a concern.[50,53] As these models progress and refine, the challenges identified in our study, such as laypeople's capacity to differentiate between AI-generated and doctors' medical responses, and biases in evaluation, will persist in their relevance and potentially even grow. It is critical for future research and development initiatives to take these insights into account when both designing AI models for incorporation into health care systems and outlining the framework for their effective, ethical implementation.

LIMITATIONS

There are several limitations to this study. First, this study uses GPT-3 rather than a more recent version of the model. While newer models might improve accuracy, it is concerning that even low-accuracy responses from older models proved convincing.

Second, our participant pool, recruited through an online platform, may be skewed toward the technologically savvy and represents mainly those 18 to 49 years of age. Additionally, participants evaluated hypothetical scenarios rather than their own medical questions, lacking personal investment in the responses. Third, the study examines single question–response pairs without the context and follow-up typical in real clinical scenarios, where doctors would likely request additional information before providing advice. Future research should explore how such context affects AI's role in medical question answering.

BROADER IMPLICATIONS

Our findings expose a number of key considerations that need to be consistently evaluated, from the perspective of both the layperson and the physician, when designing and deploying technologies such as LLMs and chatbots in medical response applications.

While existing studies of LLMs in medicine focus on physician perspectives and response accuracy,[50] our research reveals a critical disconnect: responses deemed low accuracy by physicians may still be highly persuasive to laypeople, highlighting the danger in generating and releasing AI-generated medical responses to the public without doctor supervision. This mirrors trends of patients trusting nonexpert sources for health advice[65] and confirms that people cannot reliably identify AI-generated text, which

can appear "more human than human."[66] This creates a dangerous scenario where inaccurate AI medical advice might be deemed as trustworthy as a doctor's response. Physician oversight remains essential for filtering out inaccurate or inappropriate medical information.

Both experts and laypeople show inherent biases: laypeople trust responses labeled doctor more, while experts could exhibit bias toward or against AI-generated responses. Successful deployment of medical AI systems must therefore consider both source transparency and bias mitigation. Rather than focusing solely on accuracy, stakeholders should be encouraged to use AI as an augmentation tool while maintaining critical evaluation of information.[67]

Our study reveals that participants rate AI-generated responses, particularly high-accuracy ones, as equal to or better than doctors' responses across all metrics, while maintaining higher trust in responses attributed to doctors. However, responses labeled as doctor assisted by AI showed no significant improvement in evaluations, complicating the ideal solution of combining AI's comprehensive responses with physician trust. This underscores the complexities of the situation and emphasizes the intricate dynamics through which participants and experts interact and perceive medical responses. Future doctor-assisted-by-AI applications will need careful framework design to build trust.[58] Our findings suggest three key considerations: AI can effectively deliver medical responses when accurate; inaccurate AI responses risk misleading the public through persuasive humanlike language; and expert oversight is crucial to maximize AI's unique capabilities while minimizing risks. Health care providers should focus on developing best practices for AI integration, establishing transparent physician involvement in AI-generated information delivery, and creating standardized methods for evaluating doctor–AI collaboration across medical domains.

This research reveals that AI-generated medical responses are not only indistinguishable from doctors' responses but are often preferred by the general public across all metrics — understandability, validity, trust, completeness/satisfaction, and persuasion. However, participants showed higher trust when they believed responses came from doctors. This creates a concerning paradox: while AI responses can be compelling and seemingly trustworthy, their potential inaccuracies could lead to harmful or fatal consequences if used without expert oversight. These findings suggest that integrating AI into medical information delivery requires a more nuanced approach than previously considered.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Statistical support was provided by data science specialist Steven Worthington at the Institute for Quantitative Social Science, Harvard University.

This research was reviewed and approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects, protocol number E-4170.

The raw data will be available on a GitHub repository: https://github.com/mitmedialab/LLM-MedQA.

The code for data cleaning, analysis, and visualization will be available on the same GitHub repository as the data.

Author Affiliations

[1] MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA

[2] Stanford Medicine, Stanford University, Stanford, CA

[3] Cornell Tech, New York

[4] OpenNotes, Beth Israel Deaconess Medical Center, Boston

[5] IBM Research, Yorktown Heights, NY

References

1. Beam AL, Drazen JM, Kohane IS, Leong TY, Manrai AK, Rubin EJ. Artificial intelligence in medicine. N Engl J Med 2023;388:1220-1221. DOI: 10.1056/NEJMe2206291.
2. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330:78-80. DOI: 10.1001/jama.2023.8288.
3. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med 2024;184:581-583. DOI: 10.1001/jamainternmed.2024.0295.
4. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. January 11, 2024 (https://arxiv.org/abs/2401.05654). Preprint.
5. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? Am J Emerg Med 2024;79:44-47. DOI: 10.1016/j.ajem.2024.02.008.
6. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500-510. DOI: 10.1038/s41568-018-0016-5.
7. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4:e7785. DOI: 10.2196/mental.7785.

8. Vamathevan J, Clark D, Czodrowski P, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019;18:463-477. DOI: 10.1038/s41573-019-0024-5.

9. van Kessel R, Kyriopoulos I, Wong BLH, Mossialos E. The effect of the COVID-19 pandemic on digital health–seeking behavior: big data interrupted time-series analysis of Google trends. J Med Internet Res 2023;25:e42401. DOI: 10.2196/42401.

10. Neely S, Eldredge C, Sanders R. Health information seeking behaviors on social media during the COVID-19 pandemic among American social networking site users: survey study. J Med Internet Res 2021;23:e29802. DOI: 10.2196/29802.

11. Zimmerman MS. Health information-seeking behavior in the time of COVID-19: information horizons methodology to decipher source path during a global pandemic. J Doc 2021;77:1248-1264. DOI: 10.1108/JD-01-2021-0022.

12. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: H Larochelle, M Ranzato, R Hadsell, MF Balcan, H Lin, eds. 34th Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS, 2020:1877-1901. (https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).

13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. October 11, 2018 (https://arxiv.org/abs/1810.04805). Preprint.

14. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: language models for dialog applications. January 20, 2022 (https://arxiv.org/abs/2201.08239). Preprint.

15. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. March 15, 2023 (https://arxiv.org/abs/2303.08774). Preprint.

16. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. March 22, 2023 (https://arxiv.org/abs/2303.12712). Preprint.

17. Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. February 27, 2023 (https://arxiv.org/abs/2302.13971). Preprint.

18. Taori R, Gulrajani I, Zhang T, et al. Stanford Alpaca: an instruction-following LLaMA model. 2023. Accessed April 1, 2025. https://github.com/tatsu-lab/stanford_alpaca.

19. Mehnen L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. April 27, 2023 (https://www.medrxiv.org/content/10.1101/2023.04.20.23288859v2). Preprint.

20. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023;20:3378. DOI: 10.3390/ijerph20043378.

21. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model:an observational study. The Lancet Digital Health 2024;6:E555-E561. DOI: 10.1016/S2589-7500(24)00097-9.

22. Chen S, Kann BH, Foote MB, et al. The utility of ChatGPT for cancer treatment information. March 23, 2023 (https://www.medrxiv.org/content/10.1101/2023.03.16.23287316v1). Preprint.

23. Perlis RH. Application of GPT-4 to select next-step antidepressant treatment in major depression. April 18, 2023 (https://www.medrxiv.org/content/10.1101/2023.04.14.23288595v1). Preprint.

24. He Y, Tang H, Wang D, Gu S, Ni G, Wu H. Will ChatGPT/GPT-4 be a lighthouse to guide spinal surgeons? Ann Biomed Eng 2023;51:1362-1365. DOI: 10.1007/s10439-023-03206-0.

25. Cheng K, Sun Z, He Y, Gu S, Wu H. The potential impact of ChatGPT/GPT-4 on surgery: will it topple the profession of surgeons? Int J Surg 2023;109:1545-1547. DOI: 10.1097/JS9.0000000000000388.

26. Ali H. The potential of GPT-4 as a personalized virtual assistant for bariatric surgery patients. Obes Surg 2023;33:1605. DOI: 10.1007/s11695-023-06576-5.

27. Pataranutaporn P, Danry V, Leong J, et al. AI-generated characters for supporting personalized learning and well-being. Nat Mach Intell 2021;3:1013-1022. DOI: 10.1038/s42256-021-00417-9.

28. Pataranutaporn P, Winson K, Yin P, et al. Future you: a conversation with an AI-generated future self reduces anxiety, negative emotions, and increases future self-continuity. 2024 IEEE Frontiers in Education Conference. Washington, DC: IEEE, 2024:1-10. DOI: 10.1109/FIE61694.2024.10893443.

29. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language model for mental health: a systematic review. February 19, 2024 (https://arxiv.org/abs/2403.15401). Preprint.

30. Malgaroli M, Schultebraucks K, Myrick KJ, et al. Large language models for the mental health community: framework for translating code to care. Lancet Digit Health 2025;7:e282-e285. DOI: 10.1016/S2589-7500(24)00255-3.

31. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Ment Health Res 2024;3:12. DOI: 10.1038/s44184-024-00056-z.

32. Fang CM, Liu AR, Danry V, et al. How AI and human behaviors shape psychosocial effects of chatbot use: a longitudinal randomized controlled study. March 21, 2025 (https://arxiv.org/abs/2503.17473). Preprint.

33. Liu AR, Pataranutaporn P, Maes P. Chatbot companionship: a mixed-methods study of companion chatbot usage patterns and their relationship to loneliness in active users. October 28, 2024 (https://arxiv.org/abs/2410.21596). Preprint.

34. Cadamuro J, Cabitza F, Debeljak Z, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European federation of clinical chemistry and laboratory medicine (EFLM) working group on artificial intelligence

(WG-AI). Clin Chem Lab Med 2023;61:1158-1166. DOI: 10.1515/cclm-2023-0355.

35. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023;307:230725. DOI: 10.1148/radiol.230725.

36. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus 2023;15:e37589. DOI: 10.7759/cureus.37589.

37. Liu Z, Yu X, Zhang L, et al. DeID-GPT: zero-shot medical text de-identification by GPT-4. March 20, 2023 (https://arxiv.org/abs/2303.11032). Preprint.

38. Oh N, Choi GS, Lee WY. ChatGPT goes to operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023;104:269-273. DOI: 10.4174/astr.2023.104.5.269.

39. Hisan UK, Amri MM. ChatGPT and medical education: a double-edged sword. J Pedagogy Educ Sci 2023;2:71-89. DOI: 10.56741/jpes.v2i01.302.

40. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198. DOI: 10.1371/journal.pdig.0000198.

41. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023;11:887. DOI: 10.3390/healthcare11060887.

42. Landi H. HIMSS23: Epic taps Microsoft to integrate generative AI into EHRs with Stanford, UC San Diego as early adopters. April 17, 2023. Accessed April 1, 2025. https://www.fiercehealthcare.com/health-tech/himss23-epic-taps-microsoft-integrate-generative-ai-ehrs-stanford-uc-san-diego-early.

43. Turner BE. Epic, Microsoft bring generative AI to EHRs. April 17, 2023 (https://www.modernhealthcare.com/digital-health/himss-2023-epic-microsoft-bring-openais-gpt-4-ehrs).

44. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. March 20, 2023 (https://arxiv.org/abs/2303.13375). Preprint.

45. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93:1353-1365. DOI: 10.1227/neu.0000000000002632.

46. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. April 1, 2023 (https://arxiv.org/abs/2304.01938). Preprint.

47. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP exam: a comparative study with ChatGPT-3.5. April 3, 2023 (https://www.medrxiv.org/content/10.1101/2023.04.03.23287957v1). Preprint.

48. Tanaka Y, Nakata T, Aiga K, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. PLoS Digit Health 2024;3:e0000433. DOI: 10.1371/journal.pdig.0000433.

49. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. December 26, 2022 (https://arxiv.org/abs/2212.13138). Preprint.

50. Ayers JW, Poliak A. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-596. DOI: 10.1001/jamainternmed.2023.1838.

51. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7:e2440969. DOI: 10.1001/jamanetworkopen.2024.40969.

52. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat Med 2025;31:1-6. DOI: 10.1038/s41591-024-03456-y.

53. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-1239. DOI: 10.1056/NEJMsr2214184.

54. Murphy C, Thomas FP. Generative AI in spinal cord injury research and care: opportunities and challenges ahead. J Spinal Cord Med 2023;46:341-342. DOI: 10.1080/10790268.2023.2198926.

55. Rao AS, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. J Med Internet Res 2023;25:e48659. DOI: 10.2196/48659.

56. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021:610-623. New York, NY: Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

57. Kim Y, Jeong H, Chen S, et al. Medical hallucination in foundation models and their impact on healthcare. March 3, 2025 (https://www.medrxiv.org/content/10.1101/2025.02.28.25323115v1). Preprint.

58. Jacobs M, He J, Pradier MF, et al. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In: Kitamura Y, Quigley A, eds. CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery, 2021:1-14.

59. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930-1940. DOI: 10.1038/s41591-023-02448-8.

60. Hutto C, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the Eighth International Association for the Advancement of Artificial Intelligence Conference on Web and Social Media. Copenhagen, Denmark: AAAI, 2014:216-225. DOI: 10.1609/icwsm.v8i1.14550.

61. Wrigley Kels N, Murray K, McCarthy C, O'Shea D. An objective analysis of quality and readability of online information on

COVID-19. Health Technol 2021;11:1093-1099. DOI: 10.1007/s12553-021-00574-2.

62. Szmuda T, Özdemir C, Ali S, Singh A, Syed MT, Słoniewski P. Readability of online patient education material for the novel coronavirus disease (COVID-19): a cross-sectional health literacy study. Public Health 2020;185:21-25. DOI: 10.1016/j.puhe.2020.05.041.

63. Maliha G, Gerke S, Cohen IG, Parikh RB. Artificial intelligence and liability in medicine. Milbank Q 2021;99:629-647. DOI: 10.1111/1468-0009.12504.

64. Kahr PK, Rooks G, Willemsen MC, Snijders CC. It seems smart, but it acts stupid: development of trust in AI advice in a repeated legal decision-making task. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. Sydney, NSW: IUI, 2023:528-539. DOI: 10.1145/3581641.3584058.

65. Edelman. 2023 Edelman trust barometer special report: trust and health. New York, NY: Edelman, 2023 (https://www.edelman.com/trust/2023/trust-barometer/special-report-health).

66. Jakesch M, Hancock JT, Naaman M. Human heuristics for AI-generated language are flawed. Proc Natl Acad Sci U S A 2023;120:e2208839120. DOI: 10.1073/pnas.2208839120.

67. Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In: Nicols J, ed. Proceedings of the Association for Computing Machinery on Human-Computer Interaction. Virtual event: ACM, 2021:1-21. DOI: 10.1145/3449287.