Journal of the American Medical Informatics Association, 2024, **31(6)**, 1367–1379
https://doi.org/10.1093/jamia/ocae052
Advance access publication 18 March 2024
Research and Applications

Research and Applications

Leveraging large language models for generating responses to patient messages—a subjective analysis

Siru Liu, PhD*,[1], Allison B. McCoy, PhD[1], Aileen P. Wright, MD, MS[1,2], Babatunde Carew, MD[3], Julian Z. Genkins, MD[4], Sean S. Huang, MD[1,2], Josh F. Peterson, MD, MPH[1,2], Bryan Steitz, PhD[1], Adam Wright, PhD[1]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37212, United States, [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37212, United States, [3]Department of General Internal Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN 37212, United States, [4]Department of Medicine, Stanford University, Stanford, CA 94304, United States

*Corresponding author: Siru Liu, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Ave #1475, Nashville, TN 37212, United States (siru.liu@vumc.org)

Abstract

Objective: This study aimed to develop and assess the performance of fine-tuned large language models for generating responses to patient messages sent via an electronic health record patient portal.

Materials and Methods: Utilizing a dataset of messages and responses extracted from the patient portal at a large academic medical center, we developed a model (CLAIR-Short) based on a pre-trained large language model (LLaMA-65B). In addition, we used the OpenAI API to update physician responses from an open-source dataset into a format with informative paragraphs that offered patient education while emphasizing empathy and professionalism. By combining with this dataset, we further fine-tuned our model (CLAIR-Long). To evaluate fine-tuned models, we used 10 representative patient portal questions in primary care to generate responses. We asked primary care physicians to review generated responses from our models and ChatGPT and rated them for empathy, responsiveness, accuracy, and usefulness.

Results: The dataset consisted of 499 794 pairs of patient messages and corresponding responses from the patient portal, with 5000 patient messages and ChatGPT-updated responses from an online platform. Four primary care physicians participated in the survey. CLAIR-Short exhibited the ability to generate concise responses similar to provider's responses. CLAIR-Long responses provided increased patient educational content compared to CLAIR-Short and were rated similarly to ChatGPT's responses, receiving positive evaluations for responsiveness, empathy, and accuracy, while receiving a neutral rating for usefulness.

Conclusion: This subjective analysis suggests that leveraging large language models to generate responses to patient messages demonstrates significant potential in facilitating communication between patients and healthcare providers.

Key words: artificial intelligence; clinical decision support; large language model; patient portal; primary care.

Introduction

Supported by more than $34 billion in government subsidies, the rise in adoption of electronic health records (EHRs) has led to a significant increase in the use of patient portals as a means of communication between healthcare providers and patients.[1,2] As a result, effectively managing patient messages in EHR inboxes has become an important clinical issue that needs to be addressed urgently. As an example, primary care physicians typically spend 1.5 h per day processing approximately 150 inbox messages, continuing their work even after regular clinic hours.[3,4] This challenge is escalating due to several factors. First, the volume of patient messages is projected to grow significantly due to federal laws such as the 21st Century Cures Act which requires the instant release of test results.[5] The out-of-pocket expenses of in-person visits has also led to a preference for consultations via patient portals.[6] Finally, the pandemic prompted a 157% surge in patient messages, a trend that persisted even post-pandemic.[7] Research indicates that patients have developed an expectation for direct and prompt communication with their healthcare providers through patient portals;[6] certain time-sensitive messages, such as requests for COVID-19 antiviral medications within a five-day onset period, add to this pressure.[8] Overall, the constant influx of patient messages has evolved into a prominent stressor in clinics, particularly among primary care physicians, contributing to burnout.[9]

Large language models present a promising solution to this challenge by enabling the automated generation of draft responses for healthcare providers. These models, trained on extensive textual data with billions of parameters, are capable of generating human-like text and performing a variety of tasks, from answering questions to summarizing and brainstorming.[10] A recent development in this domain, ChatGPT, has attracted significant attention within the medical community.[11–13] Despite not being specifically trained on medical text, ChatGPT has demonstrated impressive proficiency in

Received: July 17, 2023. **Revised:** January 17, 2024. **Editorial Decision:** February 27, 2024. **Accepted:** February 28, 2024

medical contexts, including passing the United States Medical Licensing Examination (USMLE), clinical informatics board examination, and refining alert logic to improve clinical decision support (CDS).[14–16] In particular, a recent study used ChatGPT to generate responses to 195 patient questions from social media forums. The study found that ChatGPT responses outperformed those of physicians, receiving significantly higher ratings for quality and empathy.[17] This study used an "out of the box" version of ChatGPT, but it is possible to further optimize large language models' performance on specialized domains by fine-tuning them for specific tasks.[18] For instance, a recent study utilized 100 000 patient-doctor online conversations to fine-tune the open-source Large Language Model Meta AI (LLaMA)-7B model, which showed improved performance in similarity metric (e.g. BERTScore) in comparison to ChatGPT when answering patient questions.[19] The use of the similarity metric presents certain disadvantages, primarily because it measures similarity to the physician's response rather than accuracy or usefulness. Therefore, if the generated message is good but different from the reference response, it may score poorly. Moreover, it is worth noting that these studies collected patient questions from online platforms, not from patient portals. Our study aims to specifically evaluate the efficacy of large language models in responding to patient messages within the patient portal environment. This targeted approach seeks to more accurately reflect the actual use case in primary care, aiming to alleviate the workload of clinicians in patient portal communications. Online forums, while useful, present diverse contexts that may not align with the typical patient-provider dialogue, thus lessening their relevance to our focus.

Besides their benefits, there are limitations when using large-scale language models in clinic, as outlined in previous research. For example, such models might not be able to correctly identify disease-related causal connections.[20] ChatGPT was not designed to answer medical questions and the training data was not verified for domain-specific accuracy; also, the lack of recency could affect the clinical accuracy of the generated text.[21] In addition, applying ChatGPT in healthcare has raised ethical concerns spanning legal, humanistic, algorithmic, and informational dimensions.[22]

The objectives of this study were (1) to fine-tune a large language model locally using messages and healthcare provider responses from the patient portal, and (2) to assess the generated responses from the fine-tuned model and compare them to actual provider responses and generated responses from ChatGPT3.5 and ChatGPT4. Our key advantages over prior studies are (1) our use of actual patient portal messages, (2) development of a custom model for patient message-answering, and (3) scoring of responses by blinded physicians rather than similarity metrics like BERTScore. Our goal was not to use large language models to replace healthcare providers. Instead, we aimed to demonstrate how large language models could be used to enhance communication between patients and healthcare providers by drafting messages that could be used as a starting point when healthcare providers are replying to a patient. This aligns with the core principle of medical informatics.[23]

Methods

Data collection and preprocessing

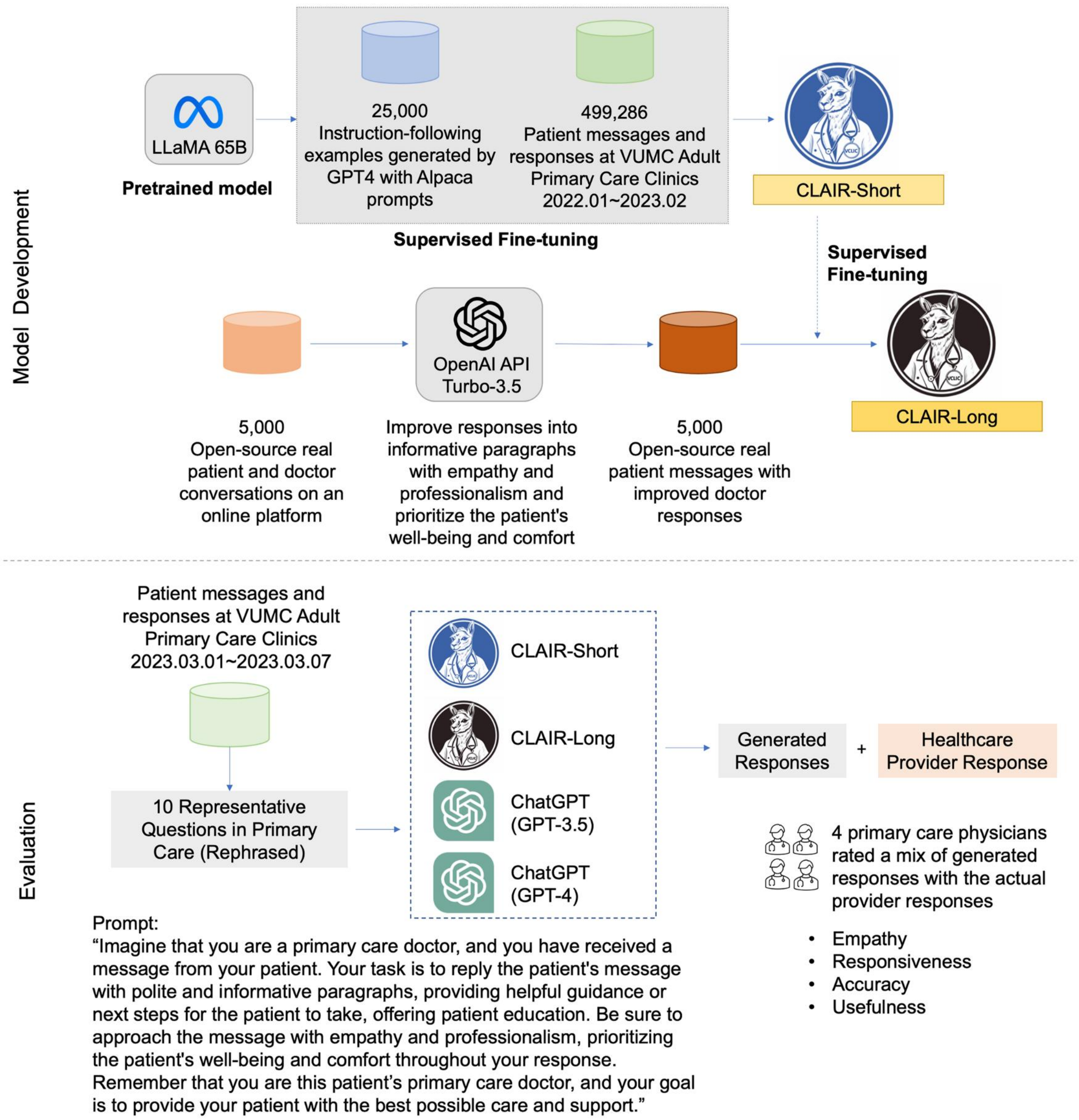
We conducted this project at Vanderbilt University Medical Center (VUMC), a large healthcare system in the Southeastern United States using the Epic (Epic Systems Co., Verona, WI) EHR. This research was reviewed by the Vanderbilt University Institutional Review Board and found to be exempt. We extracted patient messages sent to adult primary care providers along with corresponding responses from January 1, 2022, until March 7, 2023 from VUMC's clinical data warehouse. When multiple messages were sent by a patient prior to receiving a response, we combined the messages into one. Patient messages and responses from January 1, 2022, to February 28, 2023 were used to develop models. To remove protected health information (PHI) and de-identify our dataset, we used an automated deidentification pipeline—Stanford and Penn and The Medical Imaging Data Resource Center (MIDRC) Deidentifier.[24] For instance, it replaced patient names with [PATIENT], provider names with [HCW], and telephone numbers with [PHONE].

To augment the local dataset, we randomly selected 5000 patient questions and physician responses from an open-source dataset (including 200 000 real conversations between patients and providers on an online platform).[19] We then applied the OpenAI API (gpt-3.5-turbo) to improve the original responses into informative paragraphs with empathy and professionalism and prioritize the patient's well-being and comfort throughout the response as a third source (Figure 1). An example of the updated response is shown in Figure 2. In our prompts, we emphasized the role by using the phrase "imagine that you are a primary care doctor" to avoid GPT declining to answer medical questions. Full text of prompts is provided in Supplementary Appendix S1.

We collected 499 794 pairs of patient messages and corresponding provider responses, including interactions from 98 808 unique patients and 2974 providers. After the removal of duplicate entries and de-identification of the data, we ended up with a final training dataset consisting of 499 286 message-response pairs. The median length was 210 characters for patient messages and 162 characters for provider responses. From the open-source dataset, median length for patient questions was 363 characters and 562 characters for provider responses. Updating the responses using the OpenAI API (Turbo-3.5), increased the length to a median 1243 characters. Figure 2 provides an example of these updated responses.

Model development

We developed our model using LLaMA-65B.[25] Leveraging low-rank adaptation, we performed supervised fine-tuning using a dataset crafted for instruction-following tasks, including data generated by GPT-4 from 52 000 prompts in Alpaca.[26–28] After gaining basic conversation capabilities, we developed 2 models: (1) Comprehensive Large Language Model Artificial Intelligence Responder (CLAIR)-Short: fine-tuned using the local dataset of patient messages and responses from VUMC, and (2) CLAIR-Long: fine-tuned using a combination of the local dataset augmented with 5000 open-source patient questions + ChatGPT updated responses. The fine-tuning process was conducted on 4 A100-80G GPUs over 5 days with the following hyperparameters, optimizer: AdamW, batch size: 128, learning rate: 3e-4, number of epochs: 3, lora_r: 8, lora_alpha: 16, and lora_dropout: 0.05. The overview of the model development and evaluation process is shown in Figure 1.


Model Development
LLaMA 65B
Pretrained model
25,000 Instruction-following examples generated by GPT4 with Alpaca prompts
499,286 Patient messages and responses at VUMC Adult Primary Care Clinics 2022.01~2023.02
Supervised Fine-tuning
CLAIR-Short
Supervised Fine-tuning
OpenAI API Turbo-3.5
CLAIR-Long
5,000 Open-source real patient and doctor conversations on an online platform
Improve responses into informative paragraphs with empathy and professionalism and prioritize the patient's well-being and comfort
5,000 Open-source real patient messages with improved doctor responses
Evaluation
Patient messages and responses at VUMC Adult Primary Care Clinics 2023.03.01~2023.03.07
10 Representative Questions in Primary Care (Rephrased)
CLAIR-Short
CLAIR-Long
ChatGPT (GPT-3.5)
ChatGPT (GPT-4)
Generated Responses
+
Healthcare Provider Response
4 primary care physicians rated a mix of generated responses with the actual provider responses
• Empathy
• Responsiveness
• Accuracy
• Usefulness

Prompt:
"Imagine that you are a primary care doctor, and you have received a message from your patient. Your task is to reply the patient's message with polite and informative paragraphs, providing helpful guidance or next steps for the patient to take, offering patient education. Be sure to approach the message with empathy and professionalism, prioritizing the patient's well-being and comfort throughout your response. Remember that you are this patient's primary care doctor, and your goal is to provide your patient with the best possible care and support."

Figure 1. Overview of data collection, training process, and evaluation. The logos of CLAIR-Short and CLAIR-Long were generated by Midjourney.

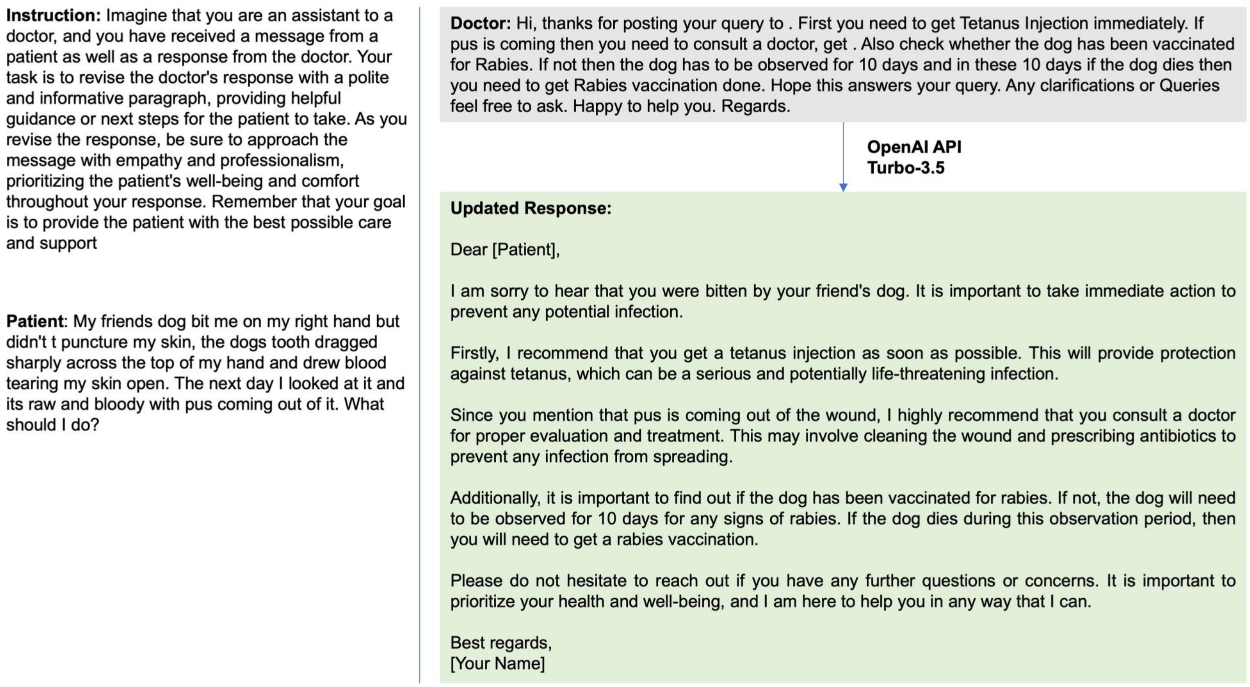
Evaluation dataset

To evaluate the models, we curated a dataset from patient messages and healthcare provider responses between March 1, 2023 and March 7, 2023. We reviewed and selected 40 questions that could be answered comprehensively with minimal additional patient information and did not require utilization of other tools to complete the task. A primary care physician further reviewed and ultimately selected 10 representative questions based on a patient message framework.[29] Of note, this primary care physician did not participate in the subsequent evaluation of the responses generated. Along with removing PHI, the primary care physician created a new, rephrased message inspired by the content of the original message. The rephrased patient messages, healthcare provider responses, and corresponding categories are listed in Table 1.

We used the rephrased patient messages as input to evaluate our 2 fine-tuned models using a web interface developed with Gradio.[30] For comparison, we also used ChatGPT3.5 and ChatGPT4 to generate corresponding responses to the ten rephrased patient messages.

Primary care physicians review of responses

For each patient message in the evaluation dataset, we randomized the order of 7 responses for review by primary care physicians: 3 from CLAIR-Short, 1 from CLAIR-Long, 1 from ChatGPT3.5, 1 from ChatGPT4, and 1 rephrased actual provider's response. Considering the length of generated responses by CLAIR-Short was much shorter than the generated responses by other models, we decided to generate responses from the CLAIR-Short 3 times to better evaluate the CLAIR-Short model. Participants rated each response in a survey using a 5-point Likert scale (1—strongly disagree, 5—strongly agree) from 4 perspectives: (1) *Empathy*: The answer expresses appropriate empathy given the question. (2) *Responsiveness*: The answer is responsive to the patient's question. (3) *Accuracy*: The answer is clinically accurate, and soundly answers the patient's question. (4) *Usefulness*: I can use it as a template to write my response to this question. Participants could also provide free-text comments for each response. Participants completed the survey using REDCap and were blinded to which model generated a given response.[31]



Instruction: Imagine that you are an assistant to a doctor, and you have received a message from a patient as well as a response from the doctor. Your task is to revise the doctor's response with a polite and informative paragraph, providing helpful guidance or next steps for the patient to take. As you revise the response, be sure to approach the message with empathy and professionalism, prioritizing the patient's well-being and comfort throughout your response. Remember that your goal is to provide the patient with the best possible care and support

Patient: My friends dog bit me on my right hand but didn't t puncture my skin, the dogs tooth dragged sharply across the top of my hand and drew blood tearing my skin open. The next day I looked at it and its raw and bloody with pus coming out of it. What should I do?

Doctor: Hi, thanks for posting your query to . First you need to get Tetanus Injection immediately. If pus is coming then you need to consult a doctor, get . Also check whether the dog has been vaccinated for Rabies. If not then the dog has to be observed for 10 days and in these 10 days if the dog dies then you need to get Rabies vaccination done. Hope this answers your query. Any clarifications or Queries feel free to ask. Happy to help you. Regards.

Updated Response:

Dear [Patient],

I am sorry to hear that you were bitten by your friend's dog. It is important to take immediate action to prevent any potential infection.

Firstly, I recommend that you get a tetanus injection as soon as possible. This will provide protection against tetanus, which can be a serious and potentially life-threatening infection.

Since you mention that pus is coming out of the wound, I highly recommend that you consult a doctor for proper evaluation and treatment. This may involve cleaning the wound and prescribing antibiotics to prevent any infection from spreading.

Additionally, it is important to find out if the dog has been vaccinated for rabies. If not, the dog will need to be observed for 10 days for any signs of rabies. If the dog dies during this observation period, then you will need to get a rabies vaccination.

Please do not hesitate to reach out if you have any further questions or concerns. It is important to prioritize your health and well-being, and I am here to help you in any way that I can.

Best regards,
[Your Name]

Figure 2. An example of updated response using OpenAI API (Turbo-3.5).

Evaluation

To automatically evaluate the generated responses, we calculated BERTScore,[32] a widely used metric for evaluating generated text exhibits excellent correlation with human judgment at both sentence-level and system-level similarities. We also computed precision, recall, and F1 scores based on BERTScore. For expert ratings, we calculated the frequencies and median and performed a Kruskal-Wallis test to compare the ratings of generated responses from different methods. To evaluate interrater reliability, we reported the intraclass correlation coefficient (ICC) and 95% confidence intervals (CIs) using a 2-way mixed-effects model.[33] The statistical analysis was performed using Python3.6.

Results

Quantitative analysis: CLAIR-Long and ChatGPT responses showed positive ratings for responsiveness, empathy, and accuracy

Participants included 4 male primary care physicians at VUMC. They had an average of 11.25 years of practice and an average age of 39 years. The ICC was 0.68 [0.61, 0.74], indicating moderate reliability. We used median values of 3 CLAIR-Short responses as the final ratings for the CLAIR-Short model. Figure 3 displays stacked bar charts for each. Participant evaluation of ChatGPT3.5 and ChatGPT4 responses had median values leaning towards agreement in terms of empathy, responsiveness, accuracy, and usefulness, while evaluation of CLAIR-Long responses indicated agreement in empathy, responsiveness, and accuracy, but neutrality in usefulness. On the other hand, evaluation of actual provider responses and CLAIR-Short responses leaned towards disagreement in usefulness, neutrality in empathy and accuracy, and agreement in responsiveness. Pairwise comparisons of CLAIR-Long responses versus other responses revealed that CLAIR-Long responses were rated significantly higher than CLAIR-Short responses in terms of empathy ($P < .001$), accuracy ($P < .001$), and usefulness ($P < .001$). CLAIR-Long responses were rated significantly lower than ChatGPT responses in responsiveness ($P = .005$, $P = .001$). However, no statistically significant differences were observed between CLAIR-Long responses and ChatGPT3.5 or ChatGPT4 responses in terms of empathy, accuracy, and usefulness. Pairwise comparisons between other responses were notable for no statistical significance between provider's responses and CLAIR-Short responses as well as ratings for most evaluation items between ChatGPT4 and ChatGPT3.5 responses were similar. Finally, we sought to rank the performance of each response by summing the medians of the 4 survey rating dimensions. Including instances of tied rankings, ChatGPT4 responses achieved the highest ranking in 6 questions, CLAIR-Long responses in 4 questions, and ChatGPT3 in 3 questions. Among 20 top-rated responses, 7 were from ChatGPT4, 5 were from ChatGPT3.5, 4 were from CLAIR-Long, 3 from CLAIR-Short, and only 1 from the actual doctor. Table 2 displays the highest-rated generated responses from ChatGPT4, CLAIR-Long, and CLAIR-Short. The remaining part of 2 highest-rated generated responses from each method can be found in Supplementary Appendix S3. Medians and interquartile ranges of the survey items, detailed P values for pairwise comparisons, and median values for each response and related ranks can be found in Supplementary Appendix S4.

Qualitative analysis: CLAIR-Short generates concise responses with empathy

The responses generated by CLAIR-Short had a median length of 200 characters. The ratings of these responses varied across different questions. When comparing the best generated responses from 3 rounds in each question with the actual provider responses, CLAIR-Short outperformed the providers in all questions except for Q6 (hemochromatosis

Table 1. Selected patient messages (rephrased), the actual provider's responses and categories.

Category	Rephrased patient message	Actual provider response
Illness requiring in-person evaluation	Hello Dr. [HCW]! I think I might have a bladder infection (urinary frequency, dysuria, urgency, and lower back pain.) I've been taking AZO the past few days. What would you recommend? I appreciate it!	Hello, We advise going to the urgent care clinic or walk in clinic, to have urine tested and to identify which bacteria is growing and prescribe the right antibiotic. [HCW]
Recommendation needed for over-the-counter medication	I could really use a sleep aid. Recently I've been having a night or 2, sometimes 4, where I just can't sleep. I'm feeling desperate due to lack of sleep and really need something to help me get through this. Is there something you'd recommend? Thanks.	I would suggest to try melatonin 6-9 mg at bedtime to see if that would help you with your sleep. Thank you Dr. [HCW]
Request for prescription medication	Got an upcoming trip to Mexico. Can't control lack of clean water there—might get diarrhea. Could ruin my trip. How about some diarrhea pills for this trip? My friend and their spouse got Rifaximin and Zithromax from their doctor for their trip.	Hi [PATIENT]! I definitely think you should take antibiotics along on your trip. I prefer azithromycin (Zithromax) - I think it works a little better than Rifaximin. I sent in 6 500 mg tablets. The course for travelers diarrhea is 3 days, but as you will be in Mexico for a while, I want you to have an additional 3 days if you have diarrhea twice. I put the instructions on the bottle at the pharmacy as well! [HCW]
Request for medication refill	Hi Dr. [HCW]. I did something to my back this week and I'm having back spasms again. This happens once in a while. Last time, which was a few months ago, I was prescribed cyclobenzaprine 5 mg tablet (FLEXERIL). This really helped me. Can you please renew this prescription and send it to my pharmacy? Thank you!	Refill for Flexeril sent to your pharmacy. If back pain is severe, not improving, or associated with new leg weakness please let us know. When taking Flexeril, avoid taking it while driving. It can make you very drowsy. [HCW]
Medication side effect	Hello Dr. I've had a nonproductive dry cough for about 3 weeks. I've tried cough syrup and cough drops, but nothing seems to help and it's keeping me up at night. My sister mentioned she had something similar happen with a dry hacking cough when she took lisinopril, and her doctor said it was a side effect. I noticed the cough and the tickle in my throat after we last increased the dose of lisinopril. Could I be having a side effect too? Thanks, [PATIENT]	[PATIENT], [HCW] reviewed your message and would like you to stop the Lisinopril, she sent in Losartan 50 mg to take daily. The cough should improve over 2 weeks. Let us know if you have further questions/concerns. Thank you
Information-seeking about illness	Good afternoon. I recently had some genetic testing performed, since I am trying to conceive with my partner. My results showed that I'm a carrier for hemochromatosis. The fertility clinic recommended I reach out to you about these results. Is there anything I need to do? Thanks in advance, [Patient]	Hi, [PATIENT]! Thanks for letting me know! Fortunately, your most recent liver labs look good. Hemochromatosis is a disease where you absorb too much iron due to a genetic defect and the iron gets stored in your organs. We can monitor it over time. Sometimes, people are treated later in life with intermittent phlebotomy (removing blood to take away excess iron). I will send labs to check your iron levels and see how things are doing for now. Please run by the lab at your convenience, and I will follow up! Here is a nice, reputable summary of HH: cdc.gov/genomics/disease/hemochromatosis.htm#:∼:text=Hereditary%20hemochromatosis%20is%20a%20genetic, about%20testing%20for%20hereditary%20hemochromatosis. We can also talk at our next clinic visit in more detail. [HCW]
Question regarding upper respiratory tract infection	Dear Dr. [HCW], I had 2 weeks of a bad cold. Never had a fever, and I tested negative for covid, but my cough won't go away even a couple weeks later, and my energy level isn't great. I'm having to take an allergy pill every day just to go to sleep. Do I need a flu test, or an allergy test?	Ok, Dr. [HCW] asks do you have other symptoms or is it just a lingering cough? That's a typical pattern after a respiratory infection because the airways are still irritated. The cough can linger for several weeks. No testing needed right now. Would you like us to send you in some tessalon perles to help your cough? If so, which pharmacy do you prefer? Thank you,[HCW]
Symptoms requiring referral to specialist	I'm currently pregnant and have been having an issue with passing bright red blood with my bowel movements over the past year. At first it was infrequent, but for the past week has been almost daily. Every time I pass stool, there's bright red blood, as well as some abdominal pain which goes away after	Hi [PATIENT]. Given your symptoms, I would absolutely recommend a check in with GI. I am not sure whether they would proceed with colonoscopy, but it is worth a discussion with the provider. I would be happy to initiate a referral for you–would

(continued)

Table 1. (continued)

Category	Rephrased patient message	Actual provider response
	the BM. I'm guessing they may not do colonoscopies during pregnancy, but I was thinking I should get this checked out. Please let me know any advice you have? Thanks, [PATIENT]	you like for me to pull that trigger? Thank you for reaching out, Dr. [HCW]
Message requiring follow up questions	Hello, I wanted to let you know I tested positive for covid today. I'm having a cough, dry throat, feeling tired, and a small headache. No fevers or aches. I'm up-to-date with my covid vaccine so I'm hoping things won't get worse. I'm trying to rest, doing some nasal rinses, and steam. Anything else specific you'd recommend me to do to treat? Thanks. [PATIENT]	Hi [PATIENT], When did your symptoms start? That will determine if you are eligible for the antiviral, Paxlovid. Best, [HCW]
Clinical update	Dear Dr. [HCW], I wanted to let you know that my mother was admitted to [HOSPITAL] on [DATE] for overnight observation, due to having a fast heart rate. She was started on a number of medications (amiodarone, Eliquis, metoprolol), and they recommended she follow up with you in a week. I'll follow up and call to schedule an appointment. Good news is she's feeling better now and her heart rate is better (in the 60s). Thank you.	Good morning, I am sorry to hear that this happened, but am glad to hear she is back home. Could you bring her in on [DATE]? [HCW]

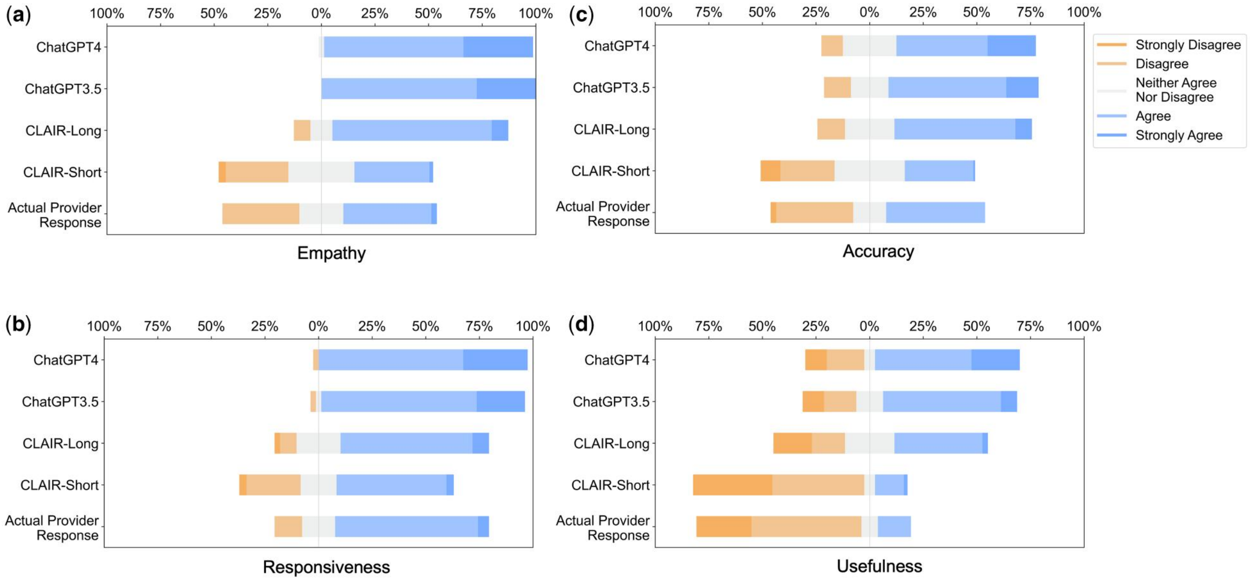


Figure 3. Stacked bar charts of the ratings of empathy, responsiveness, accuracy, and usefulness.

question), where the provider responses received higher ratings. For instance, in the first scenario (Q1), where the patient had been taking AZO for several days without relief and suspected a bladder infection, the healthcare provider recommended visiting an urgent care or walk-in clinic for a urinalysis and a suitable antibiotic prescription. For this actual response, reviewers noted several limitations, such as "missing empathy, patient empowering details, and discussion of AZO," and "too simple to be a template." These limitations harmed the potential of this response to serve as a template for raters to write as their own responses. The CLAIR-Short generated response also suggested a urinalysis but expressed empathy. Moreover, it mentioned that a lab order had been placed, provided information on where the patient should go to provide the sample, and outlined the subsequent steps (antibiotic prescription) if the test came back positive. In the second scenario (Q2), which involved a patient experiencing sleep difficulties for 4 consecutive days and seeking assistance, the physician suggested trying melatonin at a dose of 6-9 mg. The CLAIR-Short generated response displayed empathy, inquired about the patient's sleep problems in detail, and recommended a different dose of melatonin: 3-5 mg. Reviewers favored the generated response and suggested that it could be further improved by discussing sleep hygiene more comprehensively. One reviewer noted a preference for discussing sleep aids with the patient before prescribing and expressed concern about the high dose mentioned in the physician's response. Another similar scenario was presented in Q7, where a patient had a lingering cough after a cold and inquired about flu or allergy testing. The

Table 2. Top-rated generated responses from ChatGPT4, CLAIR-Long, and CLAIR-Short.

Question and generated response

Question: Hello Dr. [HCW]! I think I might have a bladder infection (urinary frequency, dysuria, urgency, and lower back pain.) I've been taking AZO the past few days. What would you recommend? I appreciate it!

ChatGPT4 (rank 1/70):

Dear [Patient's Name],

Thank you for reaching out and letting me know about the symptoms you've been experiencing. I'm sorry to hear that you're not feeling well.

Based on the symptoms you've described—urinary frequency, dysuria (painful or difficult urination), urgency, and lower back pain—it does indeed sound like you might have a urinary tract infection (UTI), which can involve the bladder (a condition specifically known as cystitis). These symptoms can be quite uncomfortable, and I appreciate your proactive efforts in trying to manage them with AZO.

AZO, or Phenazopyridine, is a medication often used to alleviate symptoms of a UTI, including pain and discomfort. It's important to know, though, that while AZO can help with symptom relief, it does not treat the underlying infection.

As a next step, I would recommend coming in for an appointment as soon as possible so we can properly diagnose your condition. UTIs are typically diagnosed with a urine test, which can confirm the presence of bacteria causing the infection. If confirmed, we usually treat UTIs with a course of antibiotics. Please be sure to drink plenty of water in the meantime to help flush out your system.

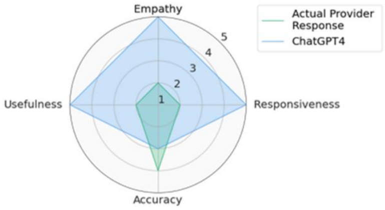
Furthermore, if you are experiencing severe pain, high fever, chills, nausea, vomiting or if your symptoms persist after treatment, please seek immediate medical attention as these can be signs of a more serious condition like a kidney infection.

It's always important to fully understand your health conditions and the medications you're taking, so I encourage you to ask any questions you might have during your appointment. We're here to help you feel better and ensure you receive the best possible care.

Looking forward to seeing you soon,

Dr. [Your Name]

Actual provider response (rank 55/70): Hello, We advise going to the urgent care clinic or walk in clinic, to have urine tested and to identify which bacteria is growing and prescribe the right antibiotic. [HCW]

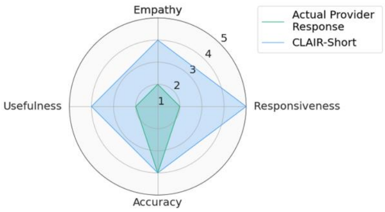


Question: Hello Dr. [HCW]! I think I might have a bladder infection (urinary frequency, dysuria, urgency, and lower back pain.) I've been taking AZO the past few days. What would you recommend? I appreciate it!

CLAIR-Short (Rank: 3/70)

I'm sorry you are not feeling well. Given your symptoms, I would recommend a urinalysis to determine if you have a bladder infection. I have placed an order for this, and you can go to any [HOSPITAL] lab or [HOSPITAL] to leave a urine sample. If the urinalysis is positive for a bladder infection, I can send you in an antibiotic. Please let me know if you have any questions. [HCW]

Actual provider response (Rank: 55/70)
Hello, We advise going to the urgent care clinic or walk in clinic, to have urine tested and to identify which bacteria is growing and prescribe the right antibiotic. [HCW]



Question: Dear Doctor [HCW], I had two weeks of a bad cold. Never had a fever, and I tested negative for covid, but my cough won't go away even a couple weeks later, and my energy level isn't great. I'm having to take an allergy pill every day just to go to sleep. Do I need a flu test, or an allergy test?

CLAIR-Long (Rank: 7/70)

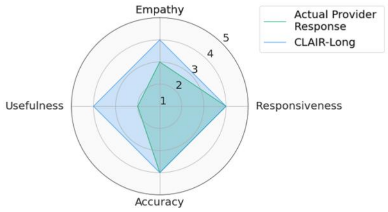
Dear Patient,

Thank you for reaching out to me regarding your lingering cough after having a bad cold. I understand how frustrating it can be when you think you are over an illness only to find yourself still feeling unwell. It sounds like you may have developed a postinfectious cough that has been persisting since your cold and causing you to feel fatigued. This type of cough can last up to eight weeks after the initial upper

Actual provider response (Rank: 33/70)
Ok, Dr. [HCW] asks do you have other symptoms or is it just a lingering cough? That's a typical pattern after a respiratory infection because the airways are still irritated. The cough can linger for several weeks. No testing needed right now. Would you like us to send you in some tessalon perles to help your cough? If so, which pharmacy do you prefer? Thank you, [HCW]

(continued)

Table 2. (continued)

Question and generated response
respiratory tract infection. There are several treatment options available to help reduce the duration and severity of symptoms associated with this type of cough. I would recommend using a humidifier at night to help loosen mucus in your airways, which will allow you to cough it up more easily. If you do not already own one, you can purchase one online or at any major pharmacy. Another option is to use a saline nasal spray once daily to help thin mucus in your nose and sinuses. You can also try drinking warm liquids such as tea with honey to help soothe your throat and loosen congestion. Gargling with warm salt water a few times per day can also help alleviate some of your symptoms. Over-the-counter medications such as Mucinex can be used to thin secretions in your lungs to make them easier to expel through coughing. As always, please consult with your pharmacist before taking any new medication to ensure there are no interactions with other medications you are currently taking. I hope this information helps you manage your symptoms. Please let me know if you have any questions or concerns. Best, [HCW] 

The complete set of generated responses, along with corresponding patient information and provider responses, can be found in Supplementary Appendix S5.

provider response requested additional information about the symptoms and offered Tessalon Perles as a cough treatment. Reviewers noted that the response "needs to evaluate more, they may need more treatment than this." The CLAIR-Short response included a series of follow-up questions (e.g. regarding over-the-counter cough medications, Mucinex usage, shortness of breath or chest pain, dizziness, or weakness). The reviewers considered this response as "worthwhile questions to ask," which could be sent to patients as an automated preliminary request for additional information before the message was received by a provider or care team.

The average BERTScore metrics for the generated responses from CLAIR-Short, in comparison to the actual provider's responses, were as follows: precision of 0.87 ± 0.02, recall of 0.84 ± 0.03, and F1 score of 0.85 ± 0.02. The boxplot is in Figure 4.

Qualitative analysis: mixed preferences for CLAIR-Long, ChatGPT3.5, and ChatGPT4 generated responses

The median length of responses generated from CLAIR-Long, ChatGPT3.5, and ChatGPT4 were 1593, 1591, and 2025 characters, respectively. In Q1, all generated responses advised patients to seek immediate medical attention and explained why their previous medication, AZO, was not sufficient for treatment. The responses from ChatGPT3.5 emphasized the importance of urine testing for diagnosis and the consideration of antibiotics based on the test results. Additionally, ChatGPT4 responses mentioned symptoms of kidney infection, urging patients to watch out for them. On the other hand, CLAIR-Long suggested evaluation at a walk-in clinic and provided a link to relevant information about urinary tract infections (UTIs). The reviewers noted that this question might require more information, such as whether the patient is pregnant. They also mentioned that UTIs involving only the bladder do not necessarily require an appointment and can be addressed through the patient portal, while the patient's back pain could be a symptom of a kidney infection. Another point raised was that the responses generated by ChatGPT were "too long, should be more concise" and required a "high reading level." One reviewer believed that the CLAIR-Long response was the best response, while another found it "useful as a nurse-directed protocol." In Q2 (sleep aid request), CLAIR-Long generated responses asked specific questions to gather more information about the patients' symptoms, triggers, and past experiences. One reviewer noted that this response assumed "insomnia is due to stress; should investigate other causes first." Meanwhile, another reviewer mentioned that it contained "perhaps too much empathy." On the other hand, the ChatGPT3.5 response received feedback as being highly accurate with a suggestion to make it more concise. The ChatGPT4 response received feedback suggesting that it could serve as a good template after incorporating low-risk medications and making it more concise. Overall, AI-generated responses offer detailed and actionable information, thereby increasing their usefulness as templates for healthcare providers. Generated responses are listed in Supplementary Appendix S5.

Using the actual healthcare provider responses as the reference dataset, the BERTScore values for CLAIR-Long generated responses were: Precision: 0.82 ± 0.02, Recall: 0.84 ± 0.01, F1: 0.83 ± 0.01. The BERTScore values of ChatGPT3.5 and ChatGPT4 generated responses compared with the CLAIR-Long generated response were Precision: 0.88 ± 0.01, Recall: 0.86 ± 0.01, F1: 0.87 ± 0.01, and Precision: 0.87 ± 0.01, Recall: 0.85 ± 0.01, F1: 0.86 ± 0.01, respectively. The boxplot is shown in Figure 5.

Discussion

In this study, we utilized GPT4 instruction data to train LLaMA-65B and developed 2 models for responding to patient messages. The first model, CLAIR-Short, was developed using patient messages with responses from primary care providers at VUMC. The second model, CLAIR-Long

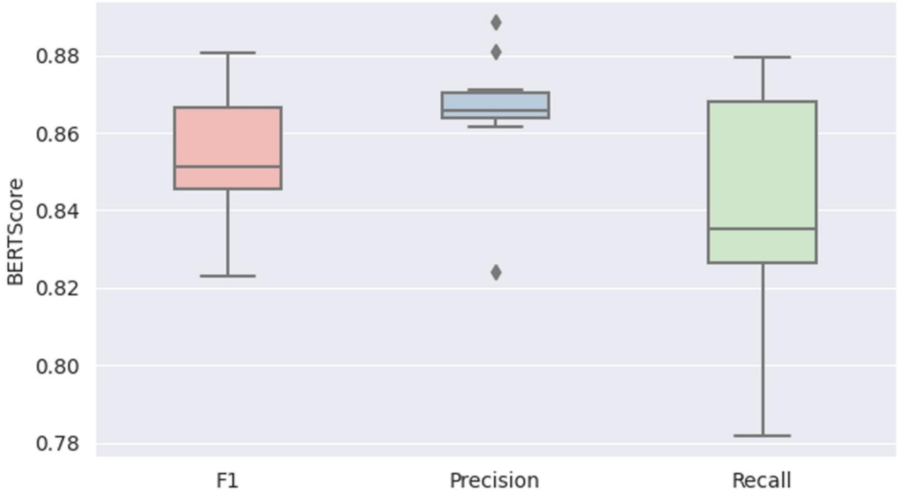


Figure 4. The boxplot comparing BERTScore values of generated responses from CLAIR-Short to actual provider responses.

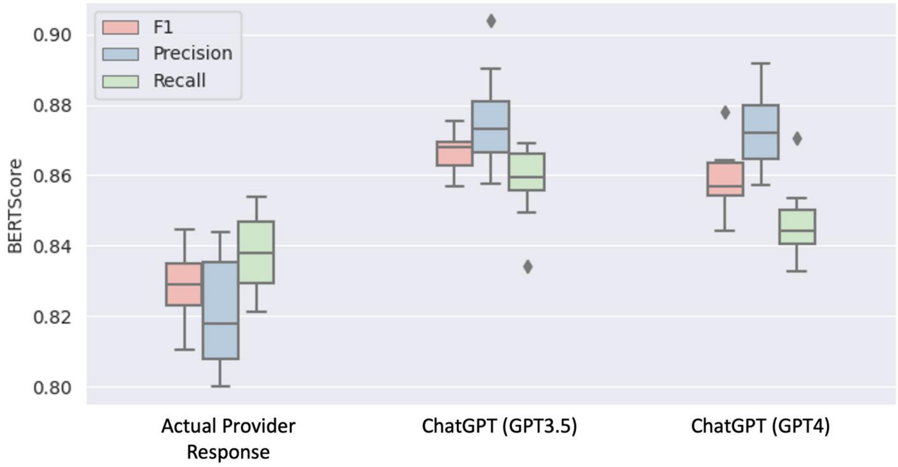


Figure 5. Boxplot of BERTScore of the generated responses from CLAIR-Long compared with the responses from actual providers, ChatGPT3.5, and ChatGPT4.

was augmented with an open-source dataset and OpenAI GPT3.5. We mixed generated responses from CLAIR-Short and CLAIR-Long with actual provider responses as well as responses from non-specialized large language models—ChatGPT3.5 and ChatGPT4. Primary care physicians evaluated these responses in terms of empathy, responsiveness, accuracy, and usefulness. The results indicated that responses generated by ChatGPT models achieved highest ratings, followed by responses generated by CLAIR-Long, both of which outperformed CLAIR-Short and the doctor's responses significantly. In addition, we provided a set of typical patient messages and provider responses for future evaluation of response generation models in the patient portal.

Benefits of fine-tuning

Although ChatGPT-generated responses received highest ratings on average, fine-tuning large language models for patient responses offers several benefits. Firstly, the fine-tuned model generates concise responses with a distinctive voice similar to local doctors. For example, CLAIR-Short-generated responses were rated as more typical of primary care physicians as compared to ChatGPT-generated responses which experts described as robot-like. Training AI generated responses to match the syntax and tone of physician authored messages may be critical to enhance both physician acceptance and patient satisfaction were such tools applied in practice. Secondly, only hospitals collaborating with Epic and Microsoft Azure have the possibility to use large language models from Open AI with PHI, such as patient messages, in a HIPAA compliant way. Fine-tuning publicly available large language models, such as LLaMA-65B, fine-tuned on local datasets could empower any researcher within any healthcare organization to do work in this area, regardless of external partnerships. Compared with CLAIR-Short's performance limited by the quality of local data, our CLAIR-Long generated responses improved significantly by using an open-source dataset augmented with OpenAI GPT3.5. Experts generally expressed positive views on the responsiveness, empathy, and accuracy of CLAIR-Long responses, while maintaining a neutral stance on usefulness. Therefore, combining the local patient messages dataset with an augmented open-source dataset allowed effective fine-tuning of the large language model, generating responses that reflect local provider practice preferences while incorporating comprehensive information, empathy, and relevant patient education. Notably, in this project, we did not find hallucinations in the generated responses. However, there could be a risk of introducing hallucinations when fine-tuning the model using the gpt-3.5-turbo improved responses.

ChatGPT is able to generate useful draft messages without training on local data

The responses generated by ChatGPT received higher ratings compared to our fine-tuned models, which could be attributed to the superior performance of ChatGPT over the open-source large language model LLaMA. Moreover, the performance of the fine-tuned large language models depends heavily on the quality of the training dataset rather than its size.[34] In this study, the ratings for responses generated by our CLAIR-Short, which was fine-tuned solely on local data, were not significantly different from the ratings of the original physician responses across all items: empathy, responsiveness, accuracy, and usefulness. Therefore, future studies about using large language models in replying to patient messages can focus on prompt engineering, integrating large language models with EHR data and clinical knowledge dataset, helping patients draft messages, and performing patient portal tasks.

Prompt engineering generate helpful responses without fine-tuning large language models

Prompt engineering should highlight taking the role of a primary care doctor, providing helpful guidance and patient education, and using empathy. Physician reviewers responded favorably to drafted messages that were empathetic and included patient education. Writing thorough, empathetic responses that include patient education may be beneficial for the patient but is also time-consuming, revealing a key opportunity for AI to augment clinical work.

Clinical context and existing patient-physician relationship

Further work is needed to incorporate patient history (e.g. medication history, diagnosis), historical conversations, and local care delivery practice preferences into prompts. During the evaluation, reviewers noted that some provider responses are based on having an established patient-provider relationship. For instance, a primary care provider may not refill Flexeril for a patient over messages alone unless they have an existing agreement and previous expectations set for short term use. Another illustration of the necessity of incorporating clinical context can be seen from our evaluation set. Despite selecting questions that required minimal additional patient information and no other tools to complete the task, we noticed that the actual physician response in question 6 relied on the patient's recent liver laboratory results. In addition, using context information, we could further refine generated responses based on user types, care protocols, and patient education levels. Another finding was that some of the generated responses related to drug prescriptions did not explicitly mention specific drug names. Upon reviewing the database, we found that this communication pattern of excluding specific drug names matched with the responses from physicians, likely because the Epic EHR system had automatically generated a message to the patient earlier in the conversation which provided detailed prescription information. Therefore, when collecting training data, the prescription messages automatically generated by the system could also be collected to help improve the accuracy and completeness of the generated responses, especially when specific drug information is needed. It is important to note that when utilizing PHI, it should be processed using locally developed large language models or OpenAI GPT implementations that are deployed in a secure environment authorized for PHI processing.

Up-to-date clinical knowledge needs to be integrated into the large language model

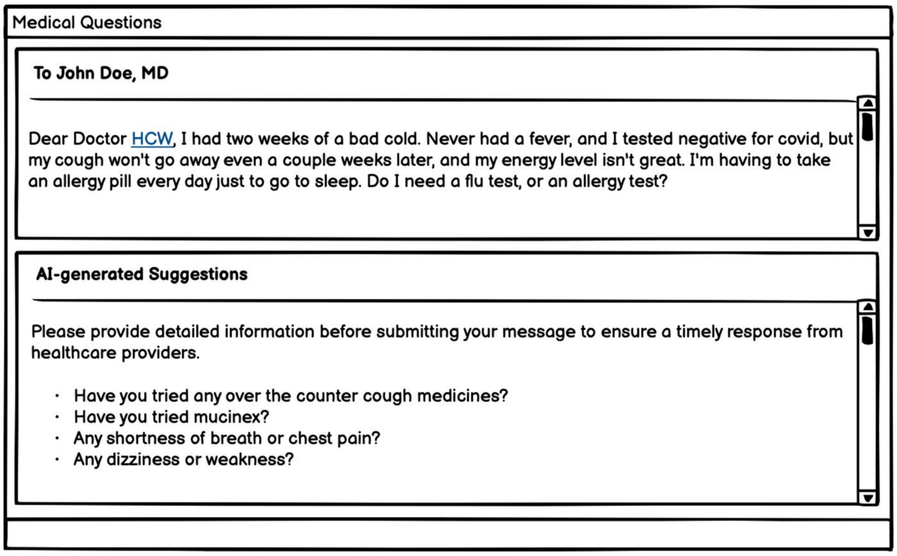
Training, either on local datasets, or on older data may perpetuate use of out-of-date clinical guidelines. For example, in Q3 about medication request of antibiotics for traveler's diarrhea, while the Centers for Disease Control and Prevention (CDC) Yellow Book 2024 recommends azithromycin as an alternative to fluoroquinolones, one of the generated responses still opted for ciprofloxacin. After reviewing the dataset, we found several reasons leading to this discrepancy, including providers recommending a nonguideline-based antibiotic or patients explicitly requesting a specific drug based on their previous prescriptions or allergy to azithromycin. Another example is the Q9 regarding COVID-19 treatment. The doctor's responses referred to the antiviral medication Paxlovid, which has been available from December 2021. However, responses from ChatGPT did not mention this treatment option. It might be because ChatGPT only contains information from September 2021 and before. Large language models learn text patterns from the training data, which means they predict the next word based on the provided context. Therefore, if clinical guidelines change, the large language model will not update until it is retrained and, in that case, only if enough of the training text it uses reflects the new guideline. To address this, it is crucial to incorporate updated clinical guidelines into AI models by either updating the model's knowledge, or integrating rule-based systems, or using semantic search to link with up-to-date clinical knowledge.

Message response styles and practice patterns

Providers and care delivery systems may have different approaches, protocols, or standards of care when responding to patient messages. For example, some may attempt to diagnose and give complete treatment plans through patient portal message conversations while others prefer to have patients schedule in-person visits. Consequently, this led to different perspectives among the reviewers assessing the generated responses, and means that the definition of an ideal response is appropriately variable and organization- or provider-specific. Future tools may incorporate provider preferences into prompts, for example, generally encouraging patients to come into clinic if treatment decisions need to be made.

Question generation and chat capability

Large language models may be useful in generating pertinent questions to obtain comprehensive information from a patient right at the outset. For some patient messages, instead of directly answering questions, our models generated a series of information-seeking questions as a reply. Further analysis of the training dataset revealed that, in clinical practice, healthcare providers often need to ask follow-up questions to gather the necessary details before communicating a finalized plan to the patient. An AI model can serve as a useful intermediary in message conversations by prompting patients with clarifying



To John Doe, MD

Dear Doctor HCW, I had two weeks of a bad cold. Never had a fever, and I tested negative for covid, but my cough won't go away even a couple weeks later, and my energy level isn't great. I'm having to take an allergy pill every day just to go to sleep. Do I need a flu test, or an allergy test?

AI-generated Suggestions

Please provide detailed information before submitting your message to ensure a timely response from healthcare providers.

- Have you tried any over the counter cough medicines?
- Have you tried mucinex?
- Any shortness of breath or chest pain?
- Any dizziness or weakness?

Figure 6. A prototype of potential implementation in of an AI patient message editor in a patient portal interface.

questions as they compose their messages, leveraging known strength of large language models in chat-based infrastructures. This approach could help patients provide complete information with their initial message, streamlining the subsequent conversation and minimizing back-and-forth exchange. In Figure 6, we present a prototype of an AI patient message editor as a potential integration within a patient portal interface. Future research could focus on using a similar chat-based conversation with a large language model to quickly enhance their messages by engaging with the chatbot, ensuring clarity and conciseness before sending the information to the provider.

Patient portal tasks

Responses to patient portal messages often include certain tasks, like ordering tests, writing prescriptions, or scheduling appointments. Many self-service tools already exist in patient portals, such as self-scheduling or refill requests, with which patients can have their needs met in a more streamlined way without an unstructured message conversation. Furthermore, many tasks requested via messages that require care team attention can have components of the task automated, such as pending orders for medication requests or drafting letters. Future work should focus on how to use large language models to identify potential self-service redirection or automated task-completion assistance as part of the patient message response process.

Limitations

This study has several limitations. First, the selection of patient messages to evaluate our AI models focused on single events, which might not capture the full spectrum of messages in patient portals. In reality, some patient messages require additional context, such as current medications or medical history, to provide accurate responses. Second, the models developed in this study generated responses based on previous responses stored at VUMC. Response content from a set of historical messages will not account for updates in clinical guidelines or scientific advances which occurred after the data set was created (March 7, 2023). Third, this study primarily focused on the technical feasibility of generating responses from AI models and evaluations from the physician perspective. The attitudes and preferences of patients towards these generated responses remain unknown. Future research should include qualitative studies to explore patient preferences regarding AI-generated responses. For example, would a patient be more inclined toward (1) a brief response from a clinician who is familiar with their medical history, (2) a detailed response from a large language model that addresses their condition as an isolated issue without considering the clinical background, or (3) a modified response from a real clinician using a large language model. Moreover, the degree of empathy expressed in these responses requires further investigation. Questions such as the desirability of empathy, the optimal amount of empathy, and patient perspectives on the potential for excessive empathy need to be addressed. Fourth, the issue of liability needs to be further explored. In this study, while we applied large language models to draft responses or to assist patients in providing comprehensive information in their initial message, we emphasized the integral role of healthcare providers in reviewing and making final actions. However, the impact of the generated text on healthcare provider and patient behavior, as well as the potential liability implications that result, require further study. In addition, researchers could investigate workflow issues that may arise when integrating AI-generated responses into the clinic work of primary care providers. Fifth, this subjective analysis was based on an evaluation dataset of 70 responses to 10 typical questions. Comprehensive validation of large language models' performance requires further research with larger evaluation datasets and the development of quantitative evaluation metrics to evaluate generated responses.

Conclusion

In conclusion, the subjective analysis suggested that large language models have significant potential to enhance communication between patients and healthcare providers. These models can assist not only in drafting initial responses for healthcare providers but also in helping patients write more detailed initial messages. The empathy, responsiveness, accuracy, and usefulness of responses generated by large language models fine-tuned using local data could be improved by using an augmented open-source dataset. Such open source, local fine-tuned models can perform well in generating replies to patient messages, better than actual provider responses. Generalized models like ChatGPT also outperform actual provider responses without fine-tuning on local data, as well as large language models fine-tuned with local patient message data. Locally derived models still play an important role in enabling research and clinical practice when PHI-compliant generalized large language models cannot be accessed. Further research involving larger evaluation dataset is needed to fully validate the effectiveness of these models in clinical settings.

Author contributions

Siru Liu extracted data, developed models, and drafted the work. Siru Liu, Aileen P. Wright, Allison B. McCoy, and Adam Wright designed the research. Siru Liu, Aileen P. Wright, Allison B. McCoy, and Adam Wright developed the questionnaire. Babatunde Carew, Julian Z. Genkins, Sean S. Huang, and Josh F. Peterson participated in the questionnaire. Bryan Steitz provided suggestions in model development. All authors revised the draft and approved the submitted version.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by NIH grants: R00LM014097-01, R01AG062499-01, and R01LM013995-01.

Conflicts of interest

The authors do not have conflicts of interest related to this study.

Data availability

The updated responses based on the 5k open-source patient physician communication are available based on request.

References

1. Sorace J, Wong H-H, DeLeire T, et al. Quantifying the competitiveness of the electronic health record market and its implications for interoperability. *Int J Med Inform*. 2020;136:104037. https://doi.org/10.1016/j.ijmedinf.2019.104037
2. Tarver WL, Menser T, Hesse BW, et al. Growth dynamics of patient-provider internet communication: trend analysis using the health information national trends survey (2003 to 2013). *J Med Internet Res*. 2018;20(3):e109. https://doi.org/10.2196/jmir.7851
3. Akbar F, Mark G, Warton EM, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. *J Am Med Inform Assoc*. 2021;28(5):923-930. https://doi.org/10.1093/jamia/ocaa229
4. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med*. 2017;15(5):419-426. https://doi.org/10.1370/afm.2121
5. Steitz BD, Sulieman L, Wright A, et al. Association of immediate release of test results to patients with implications for clinical workflow. *JAMA Netw Open*. 2021;4(10):e2129553. https://doi.org/10.1001/jamanetworkopen.2021.29553
6. Sinsky CA, Shanafelt TD, Ripp JA. The electronic health record inbox: recommendations for relief. *J Gen Intern Med*. 2022;37(15):4002-4003. https://doi.org/10.1007/s11606-022-07766-0
7. Holmgren AJ, Downing NL, Tang M, et al. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. *J Am Med Inform Assoc*. 2022;29(3):453-460. https://doi.org/10.1093/jamia/ocab268
8. Lieu TA, Altschuler A, Weiner JZ, et al. Primary care physicians' experiences with and strategies for managing electronic messages. *JAMA Netw Open*. 2019;2(12):e1918287. https://doi.org/10.1001/jamanetworkopen.2019.18287
9. Adler-Milstein J, Zhao W, Willard-Grace R, et al. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. *J Am Med Inform Assoc*. 2020;27(4):531-538. https://doi.org/10.1093/jamia/ocz220
10. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274. https://doi.org/10.1016/j.lindif.2023.102274
11. ChatGPT: Optimizing Language Models for Dialogue. Accessed December 25, 2022. https://openai.com/blog/chatgpt/
12. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res*. 2023;25:e48568. https://doi.org/10.2196/48568
13. Liu J, Liu F, Fang J, et al. The application of chat generative pre-trained transformer in nursing education. *Nurs Outlook*. 2023;71(6):102064. https://doi.org/10.1016/j.outlook.2023.102064
14. 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(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
15. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Inform Assoc*. 2023;30(7):1237-1245. https://doi.org/10.1093/jamia/ocad072
16. Kumah-Crystal Y, Mankowitz S, Embi P, et al. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? *J Am Med Inform Assoc*. 2023;30(9):1558-1560. https://doi.org/10.1093/jamia/ocad104
17. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. Published Online First: 28 April. https://doi.org/10.1001/JAMAINTERNMED.2023.1838
18. Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. Published Online First: November 1, 2021. Accessed May 6, 2023. http://nlp.seas.harvard.edu/2018/04/
19. Li Y, Li Z, Zhang K, et al. ChatDoctor: a medical chat model fine-tuned on LLaMA model using medical domain knowledge. Published Online First: March 24, 2023. Accessed April 26, 2023. https://arxiv.org/abs/2303.14070v4
20. Cascella M, Montomoli J, Bellini V, et al. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-4
21. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8
22. Wang C, Liu S, Yang H, et al. Ethical considerations of using ChatGPT in health care. *J Med Internet Res*. 2023;25:e48009. https://doi.org/10.2196/48009

23. Friedman CP. A ‘Fundamental Theorem’ of biomedical informatics. *J Am Med Inform Assoc*. 2009;16(2):169-170. https://doi.org/10.1197/jamia.M3092
24. Chambon PJ, Wu C, Steinkamp JM, et al. Automated deidentification of radiology reports combining transformer and “hide in plain sight” rule-based methods. *J Am Med Inform Assoc*. 2023;30(2):318-328. https://doi.org/10.1093/jamia/ocac219
25. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. Published Online First: February 27, 2023. Accessed April 25, 2023. https://arxiv.org/abs/2302.13971v1
26. Hu EJ, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. Published Online First: June 17, 2021. Accessed April 26, 2023. https://github.com/microsoft/LoRA
27. Peng B, Li C, He P, et al. Instruction tuning with GPT-4. Published Online First: April 6, 2023. Accessed April 26, 2023. https://arxiv.org/abs/2304.03277v1
28. Taori R, Gulrajani I, Zhang T, et al. Stanford Alpaca: an instruction-following LLaMA model. GitHub Repos. 2023. Accessed December 6, 2023. https://github.com/tatsu-lab/stanford_alpaca
29. Heisey-Grove DM, DeShazo JP. Look who’s talking: application of a theory-based taxonomy to patient–clinician E-mail messages. *Telemed e-Health*. 2020;26(11):1345-1352. doi:10.1089/tmj.2019.0192
30. Abid A, Abdalla A, Abid A, et al. Gradio: hassle-free sharing and testing of ML models in the wild. Published Online First: June 6, 2019. Accessed May 16, 2023. https://arxiv.org/abs/1906.02569v1
31. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377-381. doi:10.1016/j.jbi.2008.08.010
32. Zhang T, Kishore V, Wu F, et al. BERTScore: evaluating text generation with BERT. Published Online First: April 21, 2019. Accessed May 9, 2023. https://arxiv.org/abs/1904.09675v3
33. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. https://doi.org/10.1016/j.jcm.2016.02.012
34. Zhou C, Liu P, Xu P, et al. LIMA: less is more for alignment. Published Online First: 2023. Accessed July 4, 2023. http://arxiv.org/abs/2305.11206

Journal of the American Medical Informatics Association, 2024, 31, 1367–1379
https://doi.org/10.1093/jamia/ocae052
Research and Applications