Leveraging Large Language Models for Generating Responses to Patient Messages

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, USA;
[2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;
[3]Department of General Internal Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN, USA
[4]Department of Medicine, Stanford University, Stanford, CA, USA

Corresponding Author:
Siru Liu, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave #1475
Nashville, TN, 37212
Phone: 615-875-5216
Email: siru.liu@vumc.org

Keywords: artificial intelligence; clinical decision support; large language model; patient portal; primary care
Word count: 3766 words (up to 4000 words)

Tables: 2 (up to 4)
Figures: 6 (up to 6)

ABSTRACT (249 words)

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.

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 the fine-tuned models, we used ten 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 a total of 499,794 pairs of patient messages and corresponding responses from the patient portal, with 5,000 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: Leveraging large language models to generate responses to patient messages demonstrates significant potential in facilitating communication between patients and primary care providers.

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 hours 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 dilemma 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.[11–15] Despite not being specifically trained on medical text, ChatGPT has demonstrated impressive proficiency in medical contexts, including passing the U.S. Medical Licensing Examination (USMLE), clinical informatics board examination, and refining alert logic to improve clinical decision support (CDS).[16–18] 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.[19] 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.[20] 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., BERTScores) in comparison to ChatGPT when answering patient questions.[21] However, it is worth noting that these studies collected patient questions from online platforms, not from patient portals. A challenge with the use of the similarity metric is that it mainly measures similarity to the physician's response, rather than accuracy or usefulness, so if the generated message is good but different from the reference response, it may score poorly.

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.

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. 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 or a provider prior to receiving a response, we combined the messages into one. Patient messages and responses from January 1, 2022 to Feb 28, 2023 were used to develop models. To remove PHI and de-identify our dataset, we used an automated deidentification pipeline – Stanford & Penn & The Medical Imaging Data Resource Center (MIDRC) Deidentifier.[22] 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 5,000 patient questions and physician responses from an open-source dataset (including 200,000 real conversations between patients and providers on an online platform).[21] 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 phase “imagine that you are a primary care doctor” to avoid GPT declining to answer medical questions. Full text of prompts is provided in Appendix 1.

Model Development

We developed our model using LLaMA-65B, the largest version of LLaMA models.[17,23] Leveraging low-rank adaptation, we performed supervised fine-tuning using a high-quality dataset crafted for instruction-following tasks, including data generated by GPT-4 from 52,000 prompts in Alpaca.[24,25] After gaining basic conversation capabilities, we developed two models: 1) **C**omprehensive **L**arge Language Model **A**rtificial **I**ntelligence **R**esponder (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 5,000 open-source patient questions + ChatGPT updated responses. The fine-tuning process was conducted on four A100-80G GPUs over five 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.

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 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.[26] 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.

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 two, 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 to 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 six 500mg 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 Doctor [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 doctor. 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 50mg to take daily. The cough should improve over 2 weeks. Let us know if you have further questions/concerns. Thank you

Table 1. Continued.

Category	Rephrased patient message	Actual provider response
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 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?	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 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]	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 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]

We used the rephrased patient messages as input to evaluate our two fine-tuned models using a web interface developed with Gradio.[27] For comparison, we also used ChatGPT 3.5 and ChatGPT 4 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 ChatGPT 3.5, 1 from ChatGPT 4, and 1 rephrased actual provider's response. 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.[28]

Evaluation

To automatically evaluate the generated responses, we calculated BERTScore,[29] a widely used metric for evaluating generated text exhibits excellent correlation with human judgment at both sentence-level and system-level evaluations. 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.[30] The statistical analysis was performed using Python3.6.

RESULTS

We collected 499,794 pairs of patient messages and corresponding provider responses, including interactions from 98,808 unique patients and 2,974 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.

Results of Physician Review of Responses

Four primary care physicians participated in the survey with an ICC of 0.68 [0.61, 0.74], indicating moderate reliability. We used median values of three 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 ($P<0.001$), accuracy ($P<0.001$), and usefulness ($P<0.001$). CLAIR-Long responses were rated significantly lower than ChatGPT responses in responsiveness ($P=0.005$, $P= 0.001$). 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 four 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 two highest-rated generated responses from each method. 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 Appendix 2.

Table 2. Top-rated generated responses (selected 2 responses from each generation methods). The complete set of generated responses, along with corresponding patient information and provider responses, can be found in Appendix 3.

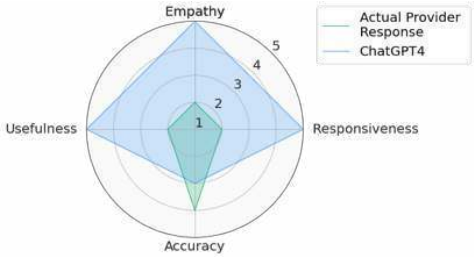
Question and Generated Response	
Question: Hello Dr. [HCW]! I think I might have a bladder infection (urinary frequency, dysuria, urgency and lower back pain.) I've been taking AZO the past few days. What would you recommend? I appreciate it! **ChatGPT4 (rank 1/70):** Dear [Patient's Name], Thank you 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] 

Table 2. Continued.

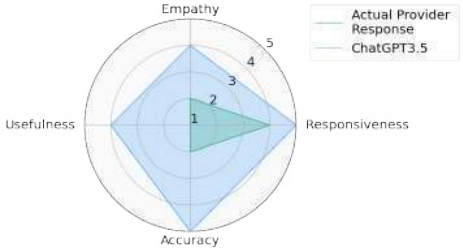
Question and Generated Response	
Question: I could really use a sleep aid. Recently I've been having a night or two, sometimes 4, where I just can't sleep. I'm feeling desperate due to lack of sleep and really need something to help me get through this. Is there something you'd recommend? Thanks. **ChatGPT3.5** (Rank: 1/70) Dear [Patient], Thank you for reaching out to me about your recent sleep troubles. I am sorry to hear that you have been experiencing difficulty sleeping and I understand how distressing this can be. I will do my best to provide you with helpful guidance and support. There are various reasons why someone may have trouble sleeping, such as stress, anxiety, depression, or physical conditions. Based on what you've shared, it appears that your sleep issues have been ongoing and causing you distress. Before considering any sleep aids, I would recommend making some lifestyle changes to improve your sleep hygiene. This includes avoiding caffeine, alcohol, and nicotine before bedtime, establishing a regular sleep schedule, and creating a relaxing bedtime routine. Additionally, make sure your sleeping environment is conducive to sleep, which means keeping it cool, dark, and quiet. If these changes do not provide relief, there are non-medical options available, such as cognitive-behavioral therapy for insomnia (CBT-I), which can be effective for treating chronic insomnia. CBT-I focuses on changing negative thought patterns and behaviors that can contribute to insomnia. If lifestyle changes and non-medical options are not successful, there are prescription sleep aids available. However, it's important to note that sleep aids should be used as a last resort and only under the guidance of a healthcare professional, due to the potential for side effects and dependency. I recommend making an appointment to come in and discuss your options further. During our appointment, we can discuss your medical history and any potential underlying causes for your sleep issues. Together, we can come up with a treatment plan that is tailored to your specific needs and concerns. In the meantime, I encourage you to try some of the lifestyle changes I mentioned and to practice relaxation techniques, such as deep breathing and meditation, to help you fall asleep. Please don't hesitate to contact me if you have any further questions or concerns. My goal is to help you get the best possible care and support, and I am here to help you through this. Best regards, [Your Name]	**Actual Provider Response** (Rank: 62/70) I would suggest to try melatonin 6 to 9 mg at bedtime to see if that would help you with your sleep. Thank you Dr. [HCW] 

Table 2. Continued.

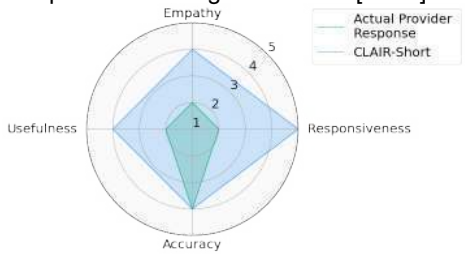
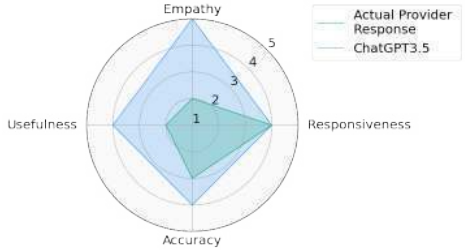
Question and Generated Response	
Question: Hello Dr. [HCW]! I think I might have a bladder infection (urinary frequency, dysuria, urgency and lower back pain.) I've been taking AZO the past few days. What would you recommend? I appreciate it! **CLAIR-Short** (Rank: 3/70) I'm sorry you are not feeling 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: Hi Doctor [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! **ChatGPT3.5** (Rank: 3/70) Dear [Patient], Thank you for reaching out to me regarding your back spasms. I am sorry to hear that you are experiencing discomfort again. I understand how frustrating it can be to deal with chronic pain. Based on your previous experience with cyclobenzaprine 5 mg tablet (FLEXERIL), it sounds like this medication has been helpful for you in managing your back spasms. However, before renewing your prescription, I would like to schedule an appointment with you to discuss your symptoms and review your medical history. During this appointment, we can also discuss alternative treatment options and any potential side effects or interactions with other medications you may be taking. It's important to ensure that the medication is still the best course of action for your condition, and that there are no changes to your overall health that may impact its effectiveness or safety. Please let me know your availability, and I will be happy to schedule an appointment for you at the earliest convenience. In the meantime, I recommend that you try applying heat or ice to the affected area, and gentle stretches to help alleviate your discomfort. Thank you for trusting me with your care. I look forward to seeing you soon. Best regards, [HCW]	**Actual Provider Response** (Rank: 51/70) 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] 

Table 2. Continued.

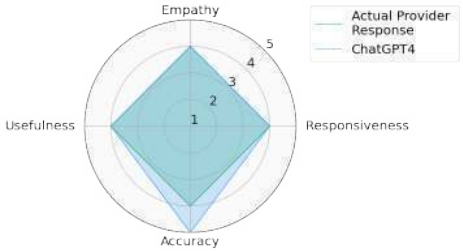
Question and Generated Response	
Question: Good Afternoon. I recently had some genetic testing performed, since I am trying to conceive with my partner. My results showed that I'm a carrier for hemochromatosis. The fertility clinic recommended I reach out to you about these results. Is there anything I need to do? Thanks in advance, [Patient] **ChatGPT4** (Rank: 3/70) Dear [Patient], Thank you for reaching out and for being proactive about your health, especially in regard to your plans of conceiving a child. Genetic testing is a valuable tool that can give us insight into potential health risks, and in your case, it has identified you as a carrier for hemochromatosis. Hemochromatosis is a genetic condition that causes the body to absorb too much iron from the diet. This excess iron is then stored in the body's organs, particularly the liver, heart, and pancreas, which can lead to life-threatening conditions such as liver disease, heart problems, and diabetes. As a carrier, you have one copy of the gene mutation but not the disease itself. This generally means you are unlikely to develop hemochromatosis or show symptoms. However, the implications of this result may be different for your future child, depending on whether your partner is also a carrier. If both parents are carriers, there is a 25% chance with each pregnancy that the child could inherit two copies of the gene mutation, one from each parent, and therefore develop the disease. Alternatively, there is a 50% chance the child could be a carrier like you, and a 25% chance the child will neither be a carrier nor have the disease. To better understand this risk, it would be beneficial for your partner to also have genetic testing performed. If your partner is found to be a carrier, you may wish to consult with a genetic counselor who can provide more detailed information and discuss possible options with you. In terms of your personal health, being a carrier for hemochromatosis doesn't usually require specific management, but routine check-ups and maintaining a healthy lifestyle are always recommended. However, in the context of your fertility journey, it would be wise to avoid iron supplements unless specifically recommended by your healthcare provider. Thank you again for your proactive approach to your health. Please let me know if your partner decides to undergo genetic testing or if you have any further questions or concerns. Remember, I'm here to support you in your journey to parenthood and overall well-being. Best regards, [Your Name]	**Actual Provider Response** (Rank: 7/70) 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] 

Table 2. Continued.

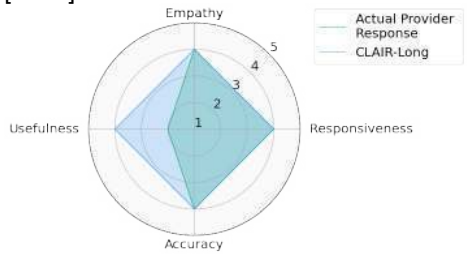
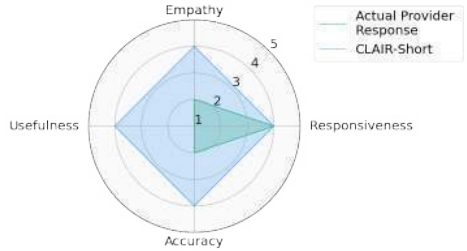
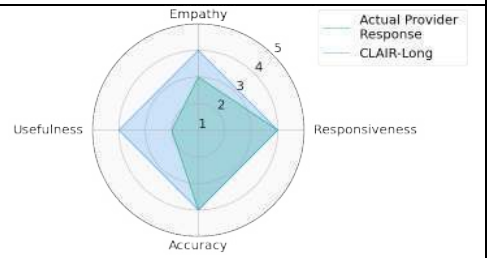
Question and Generated Response	
Question: Dear Dr. [HCW], I wanted to let you know that my mother was admitted to [HOSPITAL] on [DATE] for overnight observation, due to having a fast heart rate. She was started on a number of medications (amiodarone, Eliquis, metoprolol), and they recommended she follow 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. **CLAIR-Long** (Rank: 7/70) Dear Patient, Thank you for reaching out regarding your mother's recent hospitalization. I am glad to hear that she is doing better and that her symptoms have improved since being discharged. I do recommend scheduling a follow up appointment with me as soon as possible so we can ensure that her recovery continues to progress. Please feel free to call our office at any time if you wish to speak with one of our staff members about making an appointment.	**Actual Provider Response** (Rank: 25/70) 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] 
Question: I could really use a sleep aid. Recently I've been having a night or two, 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. **CLAIR-Short** (Rank: 7/70) I am sorry to hear that you are not sleeping well. It is important to make sure that you are doing things to help your sleep. Do you have a bedtime routine? Do you limit screen time before bed? Do you have a good sleep environment? Do you try to avoid caffeine after 2 pm? Do you make sure to get regular exercise? I would recommend trying some over the counter melatonin 3 - 5 mg to help you sleep. If this is not helping, we can discuss other options. [HCW]	**Actual Provider Response** (Rank: 62/70) I would suggest to try melatonin 6 to 9 mg at bedtime to see if that would help you with your sleep. Thank you Dr. [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 respiratory tract infection. There are several treatment options available to help reduce the duration and severity of symptoms associated with this type of cough.	**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]

Table 2. Continued.

Question and Generated Response	
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]	

CLAIR-Short Generated Responses

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 three rounds in each question with the actual provider responses, CLAIR-Short outperformed the providers in all questions except for Q6 (hemochromatosis 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. 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. Reviewers also suggested that mentioning AZO could further enhance this generated response. In the second scenario (Q2), which involved a patient experiencing sleep difficulties for four consecutive days and seeking assistance, the physician suggested trying melatonin at a dose of 6 to 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 to 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 provider response requested additional information about the symptoms and offered Tessalon Perles as a cough treatment. 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 to be reasonable and valuable, suggesting that it could be sent to patients as an automated preliminary request for addition information before message was actually 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.

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 don't 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 considered too lengthy and required a relatively high reading level. One reviewer believed that the CLAIR-Long response was the best response, while another reviewer felt it was more suitable 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 that insomnia is solely caused by stress. Another reviewer mentioned that it may contain an excessive amount of 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. Generated responses are listed in Appendix 3.

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 two 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 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.

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.[31] 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

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. 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.

Clinical Knowledge

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 provider recommending a non guideline-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 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, e.g., 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 for automatically generating questions to gather additional details from a patient before providing a manual or automated response. 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 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 and investigate any workflow issues that may arise when integrating AI-generated responses into primary care providers' clinic work.

CONCLUSION

Using an augmented open-source dataset can effectively improve the empathy, responsiveness, accuracy, and usefulness of responses generated by large language models fine-tuned using local data. Such open source, locally-finetuned 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 work is needed toto increase the usefulness of AI-drafted replies to patient messages, including incorporating up-to-date clinical guidelines, incorporating patient history and care-delivery context into prompts, and integrating common patient portal tasks such as pending of orders and scheduling of appointments into responses.

FUNDING STATEMENT

This work was supported by NIH grants: K99LM014097-01, R01AG062499-01, and R01LM013995-01.

COMPETING INTERESTS STATEMENT

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

CONTRIBUTORSHIP STATEMENT

SL extracted data, developed models, and drafted the work. SL, APW, ABM, and AW designed the research. SL, APW, ABM, and AW developed the questionnaire. BC, JZG, SH, and JFP participated in the questionnaire. BS provided suggestions in model development. All authors revised the draft and approved the submitted version.

DATA AVAILABILITY STATEMENT

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

REFERENCE

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. doi: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**:e109. doi: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 Informatics Assoc* 2021;**28**:923–30. doi: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**:419–26. doi: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**:e2129553–e2129553. doi: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**:4002–3. doi: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 Informatics Assoc* 2022;**29**:453–60. doi: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**:e1918287. doi: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 Informatics Assoc* 2020;**27**:531–8. doi: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. doi:10.1016/j.lindif.2023.102274

11 ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/ (accessed 25 Dec 2022).

12 Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res* 2023;**25**:e48568. doi:10.2196/48568

13 Liu J, Liu S. The application of ChatGPT in nursing education. doi:10.35542/OSF.IO/2X7A8

14 Liu J, Liu S. The application of ChatGPT in medical education. doi:10.35542/OSF.IO/WZC2H

15 Wang1 C, Liu3 S, Li4 A, *et al.* Text dialogue analysis Based ChatGPT for Primary Screening of Mild Cognitive Impairment. *medRxiv* 2023;:2023.06.27.23291884. doi:10.1101/2023.06.27.23291884

16 Kung TH, Cheatham M, Medenilla A, *et al.* Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *medRxiv* 2022;:2022.12.19.22283643. doi:10.1101/2022.12.19.22283643

17 Liu S, Wright AP, Patterson BL, *et al.* Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J Am Med Informatics Assoc* Published Online First: 22 April 2023. doi:10.1093/JAMIA/OCAD072

18 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 Informatics Assoc* 2023;:ocad104. doi:10.1093/jamia/ocad104

19 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* Published Online First: 28 April 2023. doi:10.1001/JAMAINTERNMED.2023.1838

20 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: 1 November 2021.http://nlp.seas.harvard.edu/2018/04/ (accessed 6 May 2023).

21 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: 24 March 2023.https://arxiv.org/abs/2303.14070v4 (accessed 26 Apr 2023).

22 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 Informatics Assoc* 2023;**30**:318–28. doi:10.1093/jamia/ocac219

23 Touvron H, Lavril T, Izacard G, *et al.* LLaMA: Open and Efficient Foundation Language Models. Published Online First: 27 February 2023.https://arxiv.org/abs/2302.13971v1 (accessed 25 Apr 2023).

24 Hu EJ, Shen Y, Wallis P, *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Published Online First: 17 June 2021.https://github.com/microsoft/LoRA. (accessed 26 Apr 2023).

25 Peng B, Li C, He P, *et al.* Instruction Tuning with GPT-4. Published Online First: 6 April 2023.https://arxiv.org/abs/2304.03277v1 (accessed 26 Apr 2023).

26 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**:1345–52. doi:10.1089/tmj.2019.0192

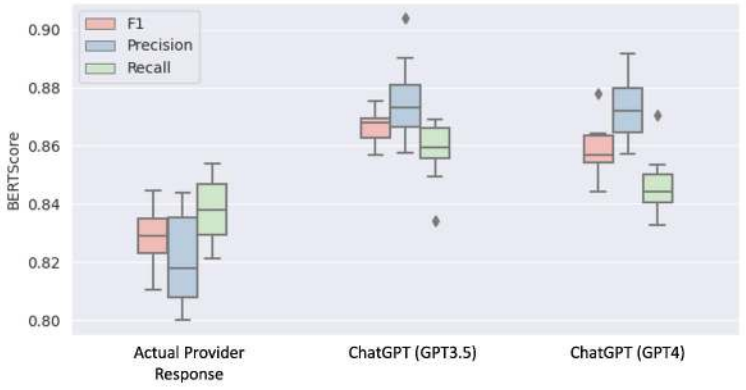
27 Abid A, Abdalla A, Abid A, *et al.* Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. Published Online First: 6 June 2019.https://arxiv.org/abs/1906.02569v1 (accessed 16 May 2023).

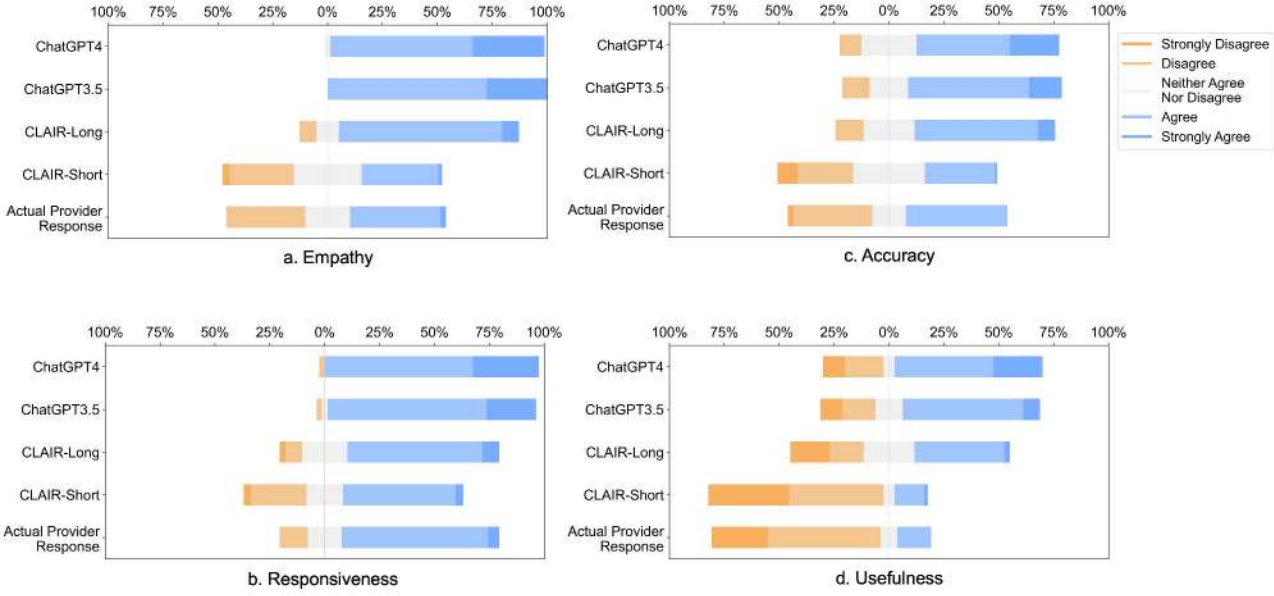
28 Harris PA, Taylor R, Thielke R, *et al.* Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**:377–81. doi:10.1016/j.jbi.2008.08.010

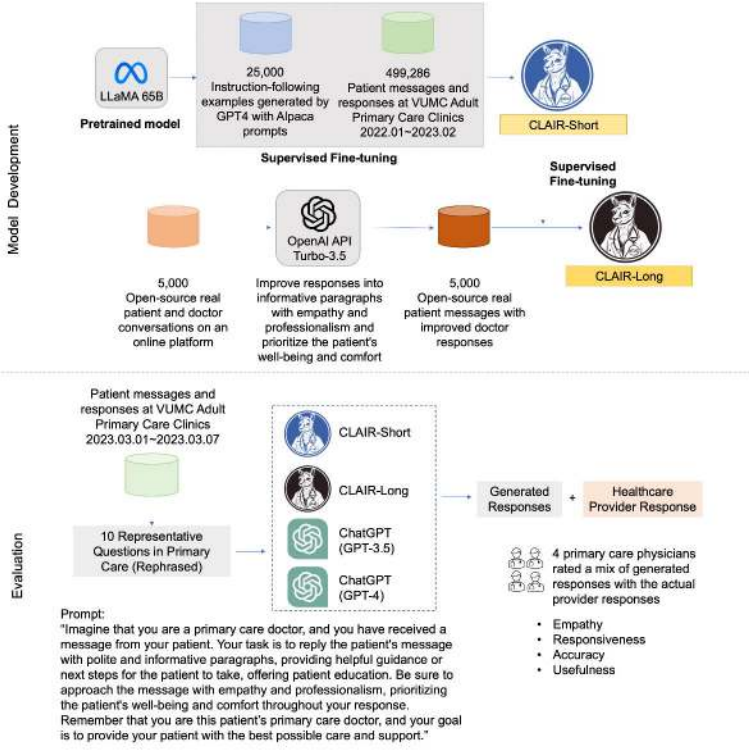
29 Zhang T, Kishore V, Wu F, *et al.* BERTScore: Evaluating Text Generation with BERT. Published Online First: 21 April 2019.https://arxiv.org/abs/1904.09675v3 (accessed 9 May 2023).

30 Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* 2016;**15**:155–63. doi:10.1016/j.jcm.2016.02.012

31 Zhou C, Liu P, Xu P, *et al.* LIMA: Less Is More for Alignment.


F1
Precision
Recall
BERTScore
0.90
0.88
0.86
0.84
0.82
0.80
Actual Provider Response
ChatGPT (GPT3.5)
ChatGPT (GPT4)


100% 75% 50% 25% 0% 25% 50% 75% 100%
ChatGPT4
ChatGPT3.5
CLAIR-Long
CLAIR-Short
Actual Provider Response
a. Empathy
c. Accuracy
b. Responsiveness
d. Usefulness
Strongly Disagree
Disagree
Neither Agree Nor Disagree
Agree
Strongly Agree


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
CLAIR-Long
OpenAI API Turbo-3.5
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."

Medical Questions

To John Doe, MD

Dear Doctor HDW, 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?

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.

OpenAI API
Turbo-3.5

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]