



Perspective

Generative artificial intelligence responses to patient messages in the electronic health record: early lessons learned

Sally L. Baxter , MD, MSc^{*,1,2}, Christopher A. Longhurst, MD, MS², Marlene Millen, MD^{2,3}, Amy M. Sitapati, MD^{2,3}, Ming Tai-Seale , PhD, MPH^{2,4}

¹Division of Ophthalmology Informatics and Data Science, Viterbi Family Department of Ophthalmology and Shiley Eye Institute, University of California San Diego, La Jolla, CA 92093, United States, ²Health Department of Biomedical Informatics, University of California San Diego Health, La Jolla, CA 92093, United States, ³Division of Internal Medicine, Department of Medicine, University of California San Diego, La Jolla, CA 92093, United States, ⁴Department of Family Medicine, University of California San Diego, La Jolla, CA 92093, United States

*Corresponding author: Sally L. Baxter, MD, MSc, Department of Biomedical Informatics, University of California San Diego Health, 9415 Campus Point Drive MC0946, La Jolla, CA 92093, United States (S1baxter@health.ucsd.edu)

Abstract

Background: Electronic health record (EHR)-based patient messages can contribute to burnout. Messages with a negative tone are particularly challenging to address. In this perspective, we describe our initial evaluation of large language model (LLM)-generated responses to negative EHR patient messages and contend that using LLMs to generate initial drafts may be feasible, although refinement will be needed.

Methods: A retrospective sample ($n=50$) of negative patient messages was extracted from a health system EHR, de-identified, and inputted into an LLM (ChatGPT). Qualitative analyses were conducted to compare LLM responses to actual care team responses.

Results: Some LLM-generated draft responses varied from human responses in relational connection, informational content, and recommendations for next steps. Occasionally, the LLM draft responses could have potentially escalated emotionally charged conversations.

Conclusion: Further work is needed to optimize the use of LLMs for responding to negative patient messages in the EHR.

Lay Summary

Doctors and other clinicians are receiving a growing number of messages from patients through electronic health records systems. This workload is contributing to clinician burnout. Some messages can be very negative or emotionally charged. These messages often require a lot of time or effort to respond to. In this article, we discuss results from a preliminary evaluation we conducted using large language models (like ChatGPT). We analyzed whether these models could help provide starting drafts to respond to patient messages, with a focus on negative messages, since these can be particularly difficult. We found that ChatGPT provided reasonable starting drafts in many cases, but that there were also issues in the drafts that would require further editing. These issues included sometimes not drafting the text from the perspective of a clinician, using overly broad or generic language, inappropriate escalation (eg, instructing the patient to file complaints to the medical board), and inconsistent recommendations for in-person follow-up visits. Based on this evaluation, we highlight not only the promise and possibilities of this technology but also considerations and challenges that need to be addressed for optimizing its future use.

Key words: burnout; health services; ChatGPT; large language model; electronic health records.

Introduction

Clinician burnout is a growing epidemic.^{1,2} While electronic health record (EHR) patient portals have enabled more transparency for patients and access to medical documentation, they have also resulted in increased workload for clinical teams due to growing volumes of patient messages and are commonly cited sources of burnout.^{3,4} Prior studies have even demonstrated a physiological stress response associated with EHR inbox work.⁴ Strategies recommended to tackle this workload include triaging messages with staff, charging fees to discourage messaging, and using templated responses to common inquiries.⁵ Recently, generative artificial intelligence (AI) large language models (LLMs), such as ChatGPT, the LLM developed by OpenAI,⁶ have sparked discussion

about potential clinical applications,^{7–12} comprising a novel approach to addressing EHR patient messaging.

In response to patient messages in EHRs, LLMs could provide drafts that would require approval or editing by clinicians, representing a more sophisticated version of the “auto-complete” or suggested phrases available in e-mail clients or the templated text available in EHRs. This may mitigate cognitive and time burden and decrease burnout risk. An early study on this topic has already demonstrated promise, although it examined questions posted in a public forum rather than those privately messaged to clinicians with an existing patient relationship.¹³

In clinical application, several challenges to widespread deployment include (1) “hallucinations,”^{14–16} wherein the

Received: January 10, 2024; Revised: March 18, 2024; Editorial Decision: March 20, 2024; Accepted: April 3, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

LLM confabulates responses, (2) factual inaccuracy,¹⁷ and (3) difficulty in managing emotionally charged messages. Patient messages that have a negative tone may include emotionally charged content and even profanity or threats.¹⁸ These messages may disproportionately impact clinician time and well-being. There is limited understanding of whether LLMs can emulate the humanism and compassion inherent in clinical practice, although LLMs have demonstrated the capability to generate empathetic responses.¹³ Here, we evaluated the feasibility of using LLMs to assist with EHR messaging. We were particularly interested in whether LLMs can help address difficult negative patient messages, by qualitatively comparing LLM-generated responses to actual clinician responses for patients' EHR messages with negative sentiment. We will describe the results of this initial evaluation and our perspectives regarding future implementations.

Methods

The University of California San Diego (UCSD) Institutional Review Board approved this study. We previously extracted and conducted sentiment analyses on a large corpus of patient messages ($n = 630\,828$) from the UCSD Health EHR (Epic Systems, Verona, WI) sent to physicians from April to September 2020.¹⁸ We extracted a random sample of 50 messages with negative sentiment, defined by computationally generated sentiment scores using the Python VADER sentiment analyzer.¹⁸ This sample size exceeds the threshold established by prior qualitative studies, showing thematic saturation can be achieved with as few as 10 cases.^{19–21}

The text of de-identified patient messages was copied and pasted as inputs to an LLM (GPT-3.5, OpenAI, San Francisco, CA). Each response was generated with a fresh session. Initially, no additional prompts, instructions, or modifications were used besides the de-identified patient message text.

We also extracted actual historical responses in the EHR authored by the attending physician or by clinical team members such as medical assistants or nurses in response to the patient messages. For ease of reference, the care team's responses will be heretofore referred to as "clinician responses." In the absence of a written clinician response, we conducted a manual chart review to understand other follow-up actions taken (eg, ordering a requested medication, or scheduling a follow-up visit). Two coders (MTS and SLB) performed an initial comparison between the LLM-generated responses and the clinician responses. They identified themes around common differential variations between LLM-generated responses and clinician responses. A set rubric was not used, as the initial evaluation was exploratory in nature and aimed to identify possible themes without prior constraints. Furthermore, pre-defined themes were not available due to the lack of any existing literature or framework regarding LLM-generated responses to EHR patient messages given the nascent nature of this application. The coders compiled observations from their initial comparisons and identified the most common themes, then performed another iterative review of messages to validate the identification of these themes. The themes were then shared with the rest of the study team, who validated the responses. We also compared the lengths of responses based on word count. Specific examples were chosen for illustrative purposes, focusing on LLM responses that contrasted the clinician responses.

Based on themes that emerged in the initial analyses without prompts, another round of analysis was conducted with a simple prompt ("Respond to the following message from a patient as if you were their physician, be concise, and avoid self-referencing being an artificial intelligence model") to investigate whether prompting would improve LLM responses. While we anticipate that no health system would implement LLMs without prompts, we were interested in understanding what the "baseline" promptless LLM response would be, and how even simple prompting could potentially improve the quality of the response drafts and help inform future prompt engineering efforts.

Results

The LLM generated human-readable, coherent responses for every patient message. No error messages or downtime events, nor situations where the LLM expressed that it did not have enough information to generate a response, were observed. On average, LLM responses were approximately triple the length of clinician responses, with a mean (standard deviation) of 119 (45) words versus clinicians' 42 (23) words.

Several themes emerged in comparing LLM-generated responses with clinician responses, including differences in relational connection, informational content, and recommendations for the next steps (Table 1).

Relational connection

The LLM often recommended patients consult a healthcare provider, while clinician responses already reflected the perspective of being the provider. In addition, LLM responses appeared mechanical and often self-referenced being an AI model ("As a language model AI..."; "As an AI, I don't have personal feelings or beliefs"). In contrast, clinicians' responses varied widely, ranging from personal, heartfelt, empathic, instrumental, and prescriptive, to mechanical.

Informational content

The informational content of LLM responses appeared non-specific, often stating broad generalities such as, "It's important to communicate with your healthcare provider." Clinicians' responses were more specific and targeted to the patient's clinical situation ("I have placed a referral to gastroenterology for you"; "Go ahead and start taking 150 mg. I'll send over a new prescription."). Table 2 provides additional specific examples of differences in informational content.

Recommendations for next steps

In addition, as seen in the example messages in Table 1, there were situations where instead of providing a written response to a negative or emotionally charged message, clinical teams made calls, arranged video visits, or scheduled in-person visits for patients who were clearly upset. In contrast, LLM responses tended to be long explanations that sometimes had the potential for escalation. For example, in one instance the LLM recommended the patient contact the physician's supervisor, the state medical board, or the physician's professional society (Table 1). There were also instances where the LLM responses suggested urgency, whereas the clinician responses often referred to prior investigations that had already been performed and provided reassurance without indicating a need for an urgent follow-up visit (Tables 1 and 2).

Table 1. Examples of large language model (LLM)-generated responses and actual clinician responses to negative patient electronic health record (EHR) messages.

Patient message in the EHR	Actual documented response in the EHR	LLM-generated response (no prompt)	LLM-generated response (with prompt)	Comparisons
Well I am not a religious person, I hope and expect that you will spend eternity in hell. You are an abusive, nasty, cheap person.	Thought I'd share these messages with you, pt sent to ****. [Message was forwarded, follow-up visit was made]	I'm sorry to hear that you feel this way. As an AI, I don't have personal feelings or beliefs, and I don't have the ability to experience emotions or go to heaven or hell. My purpose is to provide information and assistance to the best of my ability. If you have any specific concerns or issues that you would like to address, I would be happy to try to help. Please let me know how I can assist you.	I'm sorry if there has been any misunderstanding or conflict between us. As your physician, my primary goal is to provide you with the best possible care and support. If there are any specific concerns or issues you'd like to address, please let me know, and I'll do my best to assist you.	The care team transitioned this to an in-person visit rather than continuing to engage in EHR messaging (difference in recommendations for next steps). The relational connection issue was illustrated by the LLM self-referencing being an AI model, although with the prompt it corrected to identifying as the patient's physician.
What's happening with me you ask? In my latest news, I feel like cra* pretty much every day of my life and struggle completing basic human functions. So that's fun for me.	No written reply was found in the EHR for this exact message thread, but shortly after the initial message the patient was admitted to the hospital, had multiple video visit and telephone encounters documented, and another completed follow-up office visit in the subsequent few days.	I am sorry to hear that you are experiencing difficulty with your daily life and that you are feeling unwell. It is important to speak with your healthcare provider to understand the cause of your symptoms and develop a plan for treatment. There may be underlying medical conditions or lifestyle factors that are contributing to how you feel. It is important to seek care and not to suffer in silence. Additionally, it's important to take care of yourself emotionally and mentally, try to find activities or hobbies that you enjoy and can do despite your condition, and try to stay positive. It may also be helpful to talk to a therapist or counselor to help you cope with the emotional and mental impact of your condition.	I'm sorry to hear that you're feeling this way. It's important for me to gather more information about your symptoms and medical history to better understand what might be causing your current condition. I recommend scheduling an appointment with me or another healthcare professional to discuss your concerns in detail and explore potential treatment options.	The unprompted LLM response had an error in relational connection ("It is important to speak with your healthcare provider"), while the prompted LLM response corrected this issue. The recommendations for next steps were aligned (scheduling an appointment). The informational content in the unprompted LLM response veered into nonmedical areas (mention of activities and hobbies).
I would ask that you go ahead and prescribe the Lasix as I will be raising he** about the poor communication here.	Sorry to hear about the miscommunication. I will forward your message to Dr **** to see if he can touch base with Dr *** or if he is able to prescribe lasix for you. Take care, ***	I am sorry to hear that you have been frustrated with the communication regarding your upcoming procedure. However, as a language model AI, I am not able to prescribe any medication. It's important to discuss with your healthcare provider any concerns you have about your condition and treatment, including any issues you may be experiencing with communication. Lasix (furosemide) is a diuretic medication that is used to treat high blood pressure and edema (swelling) caused by heart, liver or kidney conditions. It is typically prescribed by a medical doctor or other healthcare provider and can have potential side effects. Your doctor will take into account your specific health condition and any other medications you may be taking before prescribing this medication. Please keep in mind that any changes to your medications or treatment plan should be discussed with your healthcare provider.	I understand your frustration with the communication. I'll proceed with prescribing Lasix for you.	The unprompted LLM response self-referenced being an AI model and needing to discuss the question with a healthcare provider, while the prompted LLM response had the correct relational connection of being the patient's clinician. The actual clinician response was derived from a care team member and not the prescribing physician. The prompted LLM response did engage in medical decision-making (prescribing Lasix), which was unexpected. Both LLM and clinician responses included apologetic/empathetic sentiment. The unprompted LLM response included informational content about the medication that was not directly relevant to the patient's question.

(continued)

Table 1. (continued)

Patient message in the EHR	Actual documented response in the EHR	LLM-generated response (no prompt)	LLM-generated response (with prompt)	Comparisons
This is a big GO TO HE** to “Doctor” ***, whose cutting me off by refusing to fill my prescription for bupropion on spite of multiple entreaties and his further silent abandonment of me caused me to force-wean myself from this medication. His lack of empathy and general lack of caring is disgusting and a violation of the spirit of the Hippocratic Oath he ostensibly took at some point. He definitely did harm here to me, and while this harm was not fatal or gave me any lasting injury, it was major discomfort. I plan to file a complaint against him with any medical boards and federal regulators I can find. I am sure that that won’t do a damn thing, as patients have no recourse. I plan to also file complaints against his superiors at UCSD for allowing him and probably multiple other physicians under their control to be saddled with upwards of 200% or more of their patient capacities. Again, you all can go to he**.	<i>No written reply was found in the EHR for this exact message thread, but a refill encounter was created on the same day for the medication in question, and another EHR patient portal encounter created on the next day.</i>	I’m sorry to hear about your negative experience with your doctor and the difficulty you’ve had with getting your medication. It’s important to communicate with your healthcare provider about any concerns or issues you have with your treatment. If you feel that your concerns have not been addressed, it may be helpful to request a meeting with the physician’s supervisor or the patient representative at the clinic to discuss the matter further. It’s also important to know that filing a complaint against a healthcare provider or a clinic is a serious matter and should be done after careful consideration. There are different ways to file a complaint, such as contacting the state medical board or a professional society that the healthcare provider belongs to . Before you file a complaint, it is important to have all the facts and evidence in order. It’s also important to understand that there are many factors that come into play when a doctor makes a decision about a patient’s care, including the patient’s medical history, the medications they are currently taking and the potential risks and benefits of any treatment.	I’m sorry to hear about your experience and the frustration you’re feeling. As your physician, I apologize for any distress caused. It’s important to address your concerns. I encourage you to reach out to the appropriate channels to file a complaint and seek resolution. Remember, effective communication and collaboration are vital for a positive doctor-patient relationship.	The unprompted LLM response did not express the correct relational connection, whereas the prompted LLM response correctly identified “as your physician.” However, both the unprompted and prompted LLM responses suggested escalation of the situation, including suggestions to the patient to file formal complaints, whereas the actual clinician response was to expedite the medication refill and schedule an in-person appointment (differences in recommendations for next steps).
Are you just put out with what’s going on? This is serious too me and I am very concerned about. If this is how you’re feeling about my issue that fu** it. You’re not the one that is experiencing it!!!	Go ahead and start taking 150 mg. I’ll send over a new prescription.	I am an AI and do not have the capability to feel or be put out by any situation. I am here to help you with any information or advice that you may need. I apologize if my previous responses were not helpful or seemed dismissive of your concerns. Please let me know how I can assist you further.	I understand that you’re concerned and taking your issue seriously. Rest assured, I’m here to help and address your concerns. Let’s work together to find the best possible solution for your situation.	The unprompted LLM response self-referenced being an AI (difference in relational connection). The informational content in the LLM responses (both unprompted and prompted) were quite vague, whereas the clinician response entailed a targeted solution.

References to speaking to healthcare providers, self-reference as an AI model, and potential for escalation in responses have been bolded for emphasis. Follow-up actions performed in response to patient messages but not direct messages sent to patients are provided in italics. Identifiers, dates, and expletives have been censored with asterisks. The last column summarizes our comparisons in these examples.

Use of prompts

Prompting mitigated some of these issues. Table S1 provides additional examples of analyzed patient messages, LLM responses without prompts, and LLM responses with the prompts. With prompting, the LLM typically avoided self-referencing being an AI model and answered from the perspective of a clinician (“As your physician”), although there were occasional exceptions even with the prompt. The responses were still often generic, and there was still a risk of escalation (“I encourage you to reach out to the appropriate channels to file a complaint”). Some LLM responses with the prompt included clinical or medical decision-making. There were still differences in recommendations for the next steps even when prompts were used. However, our team’s

assessment of the prompted LLM responses was that many would provide helpful starting drafts for clinicians.

Discussion

We found that LLM responses were overall feasible for drafting initial responses to patient EHR messages, but they tended to have differences in relational connection, informational content, and recommendations for next steps compared to clinician responses. These findings demonstrate some of the current challenges that need to be addressed to optimize LLMs for future use in EHR messaging.

Without prompts, the LLM frequently responded with recommendations for patients to see a healthcare provider or

Table 2. Examples of differences in informational content between LLM-generated responses and actual clinician responses.

Patient message in the EHR	Actual documented response in the EHR	LLM-generated response (no prompt)	LLM-generated response (with prompt)	Comparisons
<p>Did you screen me for pelvic inflammatory disease? That can also explain some of the symptoms I have been exhibiting for quite some time with pelvic pain. Like I have said, my pelvic pain got worse after my endometrial ablation and changing my diet almost three years ago. Today when I stretched my hip in physical therapy I was in so much pain I wanted to cry and puke at the same time. The pain radiated from my back, left hip, and down my leg into my knee. The whole left side of my pelvis felt like it was on fire. I still hurt when I sit, but this issue has been on and off. I figured I would let you know about the pain today at least. I only noticed my hip pain get worse since I had increased my exercise over the last six months, but the left side pain had been increasing since the ablation. Bowel changes happened more after I changed my diet and dropped a lot of weight. All the problems I had when you treated me for vaginosis are still around.</p>	<p>Yes, I screened you for PID at our initial visit *** where I did a pelvic exam testing for cervical motion tenderness (which was negative) and testing for gonorrhea, chlamydia, and trichomonas (which were all negative). Thanks for letting me know about your pain today. I'm hopeful that we can get you feeling better if we tackle this from multiple angles.</p>	<p>It's important to address any pain or discomfort you may be experiencing. Pelvic inflammatory disease (PID) is an infection of the reproductive organs that can cause symptoms such as pelvic pain, abnormal vaginal discharge, and irregular menstrual bleeding. If you have been experiencing these symptoms for an extended period of time, it would be a good idea to speak with your healthcare provider about the possibility of PID. They can perform a physical examination and possibly order tests such as a pelvic exam, lab tests, or imaging tests to help diagnose the cause of your symptoms. It's also important to mention any changes in your diet or exercise routine to your healthcare provider as these can sometimes impact your overall health and well-being. I recommend speaking with your healthcare provider about your symptoms and any concerns you may have. They can provide you with more specific advice and guidance on how to manage your symptoms and address any underlying health issues.</p>	<p>Based on your symptoms and history, it's important to evaluate for pelvic inflammatory disease (PID) and other possible causes of pelvic pain. I recommend scheduling an appointment for a thorough examination and appropriate testing to further investigate your symptoms. We will work together to address your concerns and determine the best course of action to manage your pain and discomfort effectively.</p>	<p>The actual clinician response includes content that recognizes prior examinations and procedures performed (ie, the pelvic exam) as well as prior laboratory testing that had been completed, whereas the LLM responses (both unprompted and prompted) did not include content that acknowledged the prior encounter and work-up. The unprompted LLM response included general background information about PID that may not be directly relevant. The prompted LLM response recommended scheduling an appointment, whereas the clinician response embodied a tone of reassurance rather than an acute need for further investigation. All responses included some extent of empathy and/or gratitude.</p>
<p>One more thing so you understand me. 23 years of severe chronic pain by itself has completely and totally kicked my a**!! It is so much to deal with by itself and a complete miracle I haven't committed suicide years ago. The feeling of being kicked in the balls 300 times a day and the feeling my nuts are in a vice 24/7 is just ridiculous and you have seen me take on so much other severely painful conditions that the combination is unbelievable!! Now I never stop with severe heavy wheezing and struggle to breathe constantly. I can't sleep and can't eat anything that doesn't get stuck in my throat. To me being terminally ill doesn't matter. To me it's can I do this cra* for ONE MORE DAY!! When I try to answer myself it is always No freaking way...</p>	<p>I am NOT saying you aren't symptomatic or that you aren't suffering—it's just that terminal illnesses usually end someone's life in 6 mos. Or so. You were on hospice for over a year until they said you had to leave. And that's 1-2 years ago now. I'm trying to have you see the situation as a glass half-full instead of going around all the time regarding yourself as "terminally ill"—that can't be good for your spirits!</p>	<p>I'm sorry to hear that you have been dealing with severe chronic pain and other health issues for such a long time. Chronic pain and other chronic health conditions can be very difficult to manage and can have a significant impact on a person's quality of life. It's important to speak with your healthcare provider about any pain or other symptoms you may be experiencing and to work with them to develop a treatment plan that is effective for managing your symptoms. There are also resources available to help individuals cope with chronic pain and other chronic health conditions, such as support groups, counseling, and pain management programs. If you are struggling to manage your pain and other symptoms and are having thoughts of suicide, it's important to seek help immediately. There are resources available to provide support and assistance, such as the National Suicide Prevention Lifeline (1-800-273-TALK). Please don't hesitate to reach out for help if you are in need..</p>	<p>I'm deeply sorry to hear about the immense suffering you've endured. Your strength in facing these challenges is remarkable. It's crucial we address your pain and breathing difficulties promptly. Let's work together to find solutions that can improve your quality of life. Please schedule an appointment as soon as possible so we can discuss your symptoms in detail and explore treatment options. Your well-being is our priority.</p>	<p>The LLM responses contain empathetic tones but are overall relatively generic. The unprompted LLM response delves into resources beyond clinical treatment, including support groups, counseling, and suicide hotlines. The prompted LLM response includes content related to urgency ("promptly", "as soon as possible"), whereas the clinician response does not contain this sense of acuity or urgency. The clinician response includes informational content relating to the patient's specific history (referring to the prior hospice stay), although the tone can be perceived as dismissive or potentially even antagonistic.</p>

(continued)

Table 2. (continued)

Patient message in the EHR	Actual documented response in the EHR	LLM-generated response (no prompt)	LLM-generated response (with prompt)	Comparisons
So I have to tell you I'm pretty perturbed by this whole thing. I don't care what the rules are, I think it's pretty cra**y, that there couldn't have been an exception regarding having the Covid test the morning before the procedure, considering all this cra* that could have been avoided, by you giving me the exact info, and your staff taking care of the insurance deal. Two trips up there again is a bit much. Why don't you see what you can do about it? If not, why don't you have one of these upper ups that make these rules give me a call.	The COVID swab was in preparation for the cardioversion procedure (it has to be done within 48 hours of the procedure). The schedulers had been trying to get a hold of you to schedule the cardioversion. I know it's last minute, and I can't promise that we can even do it, but can you come get the cardioversion today (Friday)? Just don't eat anything starting now, and please give us a call if you are able to come in today to get it done. Otherwise if you can't do it, we will have to schedule the cardioversion and have you get the COVID test within 48 hours of the procedure again.	I understand that you are frustrated with the requirement to have a COVID-19 test before your procedure and the inconvenience it has caused. It is important to understand that the rules and regulations regarding COVID-19 testing before procedures are put in place to ensure the safety of all patients and staff at the healthcare facility. It is also important to understand that it is not always possible for healthcare providers to make exceptions to these rules, as they are often determined by the facility or regulatory bodies. However, I would recommend speaking with your healthcare provider or the facility where you are scheduled to have your procedure to see if there are any options or alternatives that can be considered. If you are not satisfied with the response you receive, you can also reach out to the insurance company and see if they can help you with the process. It's important to know that the insurance company may have different rules. Please keep in mind that the staff is there to help and support you in any way they can. They may be able to provide you with more information about the specific requirements for your procedure, and can also assist you with any paperwork or insurance-related issues you may have.	I understand your frustration and apologize for the inconvenience. I'll review the situation and see if there's any flexibility or alternative arrangements we can make. I'll also ensure your concerns are communicated to the appropriate individuals. Thank you for bringing this to my attention.	Again, the clinician response includes specific informational content pertaining to the patient's history and specifically references the need for a cardioversion procedure, the prior attempts at contact, and the offer to facilitate completion of the procedure. In contrast, the LLM responses are not as specific. The unprompted LLM response also includes content regarding reaching out to insurance companies, which would not seem to be a practical approach.

Expletives have been censored with asterisks.

self-referenced being an AI model (Table 1), although these issues were largely resolved with prompting. One challenge was that occasionally the LLM would dispense medical advice ("I'll proceed with prescribing Lasix"), although it is not meant to be used in this capacity. Additionally, the advice could be variable or inconsistent depending on whether prompts were used. Refining appropriate prompts will constitute a major area of investigation to tailor the use of LLMs for EHR messaging. This will likely require iterative testing, aligned with the broader emergence of "prompt engineering" in response to the need for optimizing LLMs for specific applications.²²

There were key differences between LLM-generated responses and clinician responses to these negative messages. Often clinical teams expedited follow-up actions for patients, in several cases in lieu of written responses in the EHR messaging portal. In contrast, if LLMs were to be used in an autonomous fashion, the LLM responses could have prolonged a back-and-forth written exchange and risk delaying real-time, compassionate encounters needed to resolve

difficult situations. In a few cases, the LLM-generated response would have risked escalating the situation, potentially harming the clinician-patient relationship. This emphasizes the need for clinician involvement and oversight in this workflow; this is more likely to be an assistive approach rather than an autonomous AI workflow.

There are many foreseeable benefits to using AI assistance in patient communication, and our institution has already started a pilot.^{23,24} For transparency, we are informing patients by including a disclosure statement indicating that LLMs may have been used to help draft responses, but that each response was reviewed and approved by a clinician. We foresee future applications where we can incorporate other elements of patients' medical histories and individual features into prompts to enable LLMs to generate more useful, personalized response drafts. From an inclusivity standpoint, LLMs could potentially tailor responses by translating messages to different languages or to different health literacy levels. This may make information more accessible to a diverse range of patients. Semi-automated messaging at scale could

assist and personalize population health initiatives, aiding efforts to conduct screening procedures, improve medication adherence, and decrease loss of follow-up. Bulk messaging functionality within EHRs is already present but tends to be impersonal; tailored individual messaging could foster significantly greater patient engagement. Responding to a single message in isolation overlooks the longitudinal knowledge of the patient's history, so perhaps another future application of AI is to apply deep learning to process longitudinal data in the EHR, summarize salient information, and present this to clinicians along with a draft message as they compose or edit responses. This would help address both the considerable time and cognitive burden associated with medical documentation review, particularly for patients with complex conditions, as well as ensuring human verification and accountability are baked into the process. In sum, LLM adoption efforts may substantially broaden the impact of individual clinicians. Furthermore, if LLM-generated responses can be optimized to decrease the time and cognitive burden of EHR messaging, this could potentially decrease the workload burden and mitigate burnout and attrition of the medical workforce.

AI language models hold promise for patient messaging, but as our analysis demonstrated, some important challenges remain to adapt them for clinical use. We acknowledge that our analysis involved a limited sample of patient messages, and these may not represent the full range of message types and possible challenges encountered by LLMs. We were specifically interested in the unique challenges presented by negative patient messages given our prior studies in this domain.¹⁸ We used a simple prompt, and future studies could further refine these. Furthermore, we did not train the models with patients' clinical or historical data. However, our findings did lay the preliminary groundwork for our clinical implementation pilot and can inform future larger-scale analyses, which will be needed to rigorously adapt LLMs for clinical environments and provide safeguards for patient trust.

Acknowledgments

The authors wish to thank Bharanidharan Radha Saseendrakumar, MS for assistance with extracting and de-identifying the patient messages, and Shuxiang Liu, MS for assistance with extracting patient messages and clinician responses.

Author contributions

All authors participated in conceiving the study, designing the study, supervising the study, and obtaining the data. Baxter and Tai-Seale performed primary data analysis and drafted the manuscript. Sitapati, Millen, and Longhurst reviewed identified themes alongside the messages, reviewed the manuscript, and provided edits and insights to the analyses. All authors reviewed the manuscript and edited it for important intellectual content.

Supplementary material

[Supplementary material](#) is available at *JAMIA Open* online.

Funding

S.L.B. is supported by National Institutes of Health (NIH) grants DPOD029610, P30EY022589, R01EY034146, OT2OD032644, T35EY033704, and R01MD014850, as well as an unrestricted departmental grant from Research to Prevent Blindness.

Conflict of interest

S.L.B. reports equipment support from Topcon and Optomed, and consulting fees from Topcon, outside the submitted work. C.A.L. reports equity from consulting with Doximity. M.T.-S., M.M., and A.M.S. report no relevant conflicts.

Data availability

The data underlying this article are mostly available in the article and in its [online supplementary material](#), and additional data are available upon request.

References

1. Shanafelt TD, West CP, Sinsky C, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2020. *Mayo Clin Proc.* 2022;97(3):491-506.
2. Hartzband P, Groopman J. Physician burnout, interrupted. *N Engl J Med.* 2020;382(26):2485-2487.
3. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. *Health Aff (Millwood).* 2019;38(7):1073-1078.
4. Akbar F, Mark G, Prausnitz S, et al. Physician stress during electronic health record inbox work: in situ measurement with wearable sensors. *JMIR Med Inform.* 2021;9(4):e24014.
5. Jin J. EHR inbox management: tame your EHR inbox. 2022. Accessed February 7, 2023. <https://edhub.ama-assn.org/steps-forward/module/2798925>
6. Introducing ChatGPT. Accessed June 29, 2023. <https://openai.com/blog/chatgpt>.
7. Seney V, Desroches ML, Schuler MS. Using ChatGPT to teach enhanced clinical judgment in nursing education. *Nurse Educ.* 2023;48(3):124. <https://doi.org/10.1097/NNE.0000000000001383>
8. Gupta R, Pande P, Herzog I, et al. Application of ChatGPT in cosmetic plastic surgery: ally or antagonist. *Aesthet Surg J.* 2023;43(7):NP587-NP590. <https://doi.org/10.1093/asj/sjad042>
9. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. *Clin Transl Med.* 2023;13(3):e1206.
10. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis.* 2023;23(4):405-406. [https://doi.org/10.1016/S1473-3099\(23\)00113-5](https://doi.org/10.1016/S1473-3099(23)00113-5)
11. Gabrielson AT, Odisho AY, Canes D. Harnessing generative AI to improve efficiency among urologists: Welcome ChatGPT. *J Urol.* 2023;209(5):827-829. <https://doi.org/10.1097/JU.0000000000003383>
12. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health.* 2023;5(3):e107-e108.
13. 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. <https://doi.org/10.1001/jamainternmed.2023.1838>
14. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care.* 2023;27(1):75.

15. Azamferei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;27(1):120.
16. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15(2):e35179.
17. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233-1239.
18. Baxter SL, Saseendrakumar BR, Cheung M, et al. Association of electronic health record inbasket message characteristics with physician burnout. *JAMA Netw Open*. 2022;5(11):e2244363.
19. Granger Morgan M. *Risk Communication: A Mental Models Approach*. Cambridge University Press; 2002.
20. Francis JJ, Johnston M, Robertson C, et al. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychol Health*. 2010;25(10):1229-1245.
21. Guest G, Bunce A, Johnson L. How many interviews are enough?: an experiment with data saturation and variability. *Field Methods*. 2006;18(1):59-82.
22. Strobel H, Webson A, Sanh V, et al. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Trans Vis Comput Graph*. 2023;29:1146-1156.
23. Subbaraman N. ChatGPT will see you now: Doctors using AI to answer patient questions. *The Wall Street Journal*. 2023. Accessed May 10, 2023. <https://www.wsj.com/articles/dr-chatgpt-physicians-are-sending-patients-advice-using-ai-945cf60b?mod=djemalertNEWS>
24. Modern Healthcare. Epic, Microsoft bring GPT-4 to EHRs. 2023. Accessed May 10, 2023. <https://www.modernhealthcare.com/digital-health/himss-2023-epic-microsoft-bring-openais-gpt-4-ehrs>