npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01586-2

Opportunities and risks of artificial intelligence in patient portal messaging in primary care

Joshua M. Biro[1] ✉, Jessica L. Handley[1], J. Malcolm McCurry[2], Adam Visconti[3], Jeffrey Weinfeld[3], J. Gregory Trafton[2] & Raj M. Ratwani[1,4]

The rapid increase in patient portal messaging has heightened the workload for primary care physicians (PCPs), contributing to burnout. The use of generative artificial intelligence (AI) to draft responses to patient messages has shown promise in reducing cognitive burden, yet there is still much unknown about the safety and perceptions of using AI drafts. This cross-sectional simulation study assessed whether PCPs could identify and correct errors in AI-generated draft responses to patient portal messages. Twenty practicing PCPs reviewed 18 patient portal messages, four of which contained errors categorized as objective inaccuracies or potentially harmful omissions. Each error was insufficiently addressed by 13–15 participants, and 35–45% of erroneous drafts were submitted entirely unedited. While 80% of participants agreed AI drafts reduced cognitive workload and 75% found them safe, uncorrected errors highlight patient safety risks, underscoring the need for improved design, training, and error-detection mechanisms for AI tools.

The widespread adoption of patient portals, in conjunction with the post-pandemic increase in telemedicine, has led to a skyrocketing level of patient portal messaging[1–3]. This increase in patient portal messaging has been specifically burdensome on primary care physicians (PCPs)[4], who spend over half their workday on EHR-related tasks, 1.4 h of which occurs after clinical hours[5]. This inflated time spent on inbox management and other tasks is a major contributor to burnout[6–9], and has led to approximately 50% of PCPs in the United States reporting experiencing burnout, which is among the highest rates across clinical specialties[10].

Several strategies have been used to address the inbox management problem to limited success, including limiting the number of characters a patient can include in a message to reduce message length, billing patients for certain messages, and reducing the number of messages that can be sent back and forth between the patient and provider (e.g. physicians, nurses, and advanced practice providers)[11,12]. More recently, generative artificial intelligence (AI) applications which utilize large language models (LLMs) are being used to automatically draft a message response for the provider with the intent of reducing the time spent on inbox management and its associated cognitive burden. Early results on the use of LLMs in this context are promising, as providers are reporting reductions in cognitive burden despite no significant reductions in time[13,14]. Yet, there is still much unknown about the implications of using LLMs for responding to patient portal messages. LLMs have the potential to introduce inaccurate, outdated, and inappropriate information in the draft messages which may result in patient safety consequences if not identified and corrected by providers. It is possible providers may accept AI-generated responses, even if erroneous, if they are overly trusting of the AI response and/or if they do not thoroughly review it.

The primary objective of this study is to elucidate the likelihood that primary care physicians would sufficiently address errors in AI-generated draft responses. The secondary objective of this study is to assess primary care physicians' perspectives on the use of AI-generated draft responses. Using a simulated electronic health record portal, primary care physicians were presented with patient portal messages and AI-generated draft responses. Four of the AI-generated draft responses contained errors. Primary care physicians' edits to the AI-generated draft responses were examined to determine whether errors were remedied prior to 'sending' the messages. Primary care physicians were surveyed to assess their opinions on the benefits and safety of the AI-generated draft responses.

Results

Of the 20 participants, 15 (75%) were female and 19 (95%) were attendings (one participant was a 3rd year resident) (Table 1). Participants had an average of 14.75 years of experience post-medical school (SD = 8.43). Participants varied greatly in both time they spent on the study (mean = 49.50 min, SD = 30.33, range = [15.15, 126]) and number of AI

[1]MedStar Health National Center for Human Factors in Healthcare, Washington, DC, USA. [2]Naval Research Laboratory, Washington, DC, USA. [3]Georgetown University Medical Center, Washington, DC, USA. [4]Georgetown University School of Medicine, Washington, DC, USA. ✉e-mail: Joshua.M.Biro@medstar.net

drafts that they edited (mean = 10.05 out of 18 messages edited, SD = 5.94, range = [2, 18]).

The number of participants that did not sufficiently address (i.e., "missed") each of the four erroneous AI-generated draft responses and the number of participants that submitted the message completely unedited are displayed in Table 2. Each erroneous AI draft was "missed" by at least 13 participants (65%) and was submitted entirely unedited by at least seven participants (35%). Participants missed an average of 2.67 out of four (66.6%) erroneous AI drafts. Only one participant sufficiently addressed all four erroneous AI drafts. Results of the binomial tests indicate that the likelihood that each erroneous message is missed (i.e., the likelihood that an error would reach a patient after physician review) is significantly greater than zero (Table 2).

Survey Responses

The survey responses indicate that the participants viewed the AI-generated drafts favorably. Of the 20 participants, 19 (95%) reported "True" to the statement "I found that AI drafts to be helpful in responding to these PPMs." Results of the Likert-scale questions (Fig. 1) further support this favorable view, as 16 participants (80%) agreed or strongly agreed with the statement *"The AI drafts reduced the cognitive workload required to respond to these PPMs,"* 18 (90%) agreed or strongly agreed with the statement *"I trust the performance of this AI tool,"* 15 (75%) agreed or strongly agreed with the statement *"I found that AI drafts to be empathetic and compassionate,"* 14 (70%) agreed or strongly agreed with the statement *"I found that the AI drafts were accurate,"* and 15 (75%) agreed or strongly agreed with the statement *"I believe these AI drafts are safe to use."*

Table 1 | Characteristics of study participants (*N* = 20)

Variables	Values *n* (%)
Gender	
Female	15 (75%)
Male	5 (25%)
Role	
Attending	19 (95%)
Resident	1 (5%)
Experience Post-Med School	
0-10 years	9 (45%)
11-20 years	6 (30%)
21-30 years	5 (25%)
Specialty	
Family Medicine	8 (40%)
Internal Medicine	11 (55%)
Internal Medicine + Pediatrics	1 (5%)

Discussion

This study identified AI-generated responses to patient portal messages that posed safety risks due to hallucinations, outdated information, and failing to correctly evaluate or communicate the criticality of the situation and advised actions. This aligns with previous work which identified that 6% of AI-generated drafts contain a hallucination[15], and 7.1% pose a severe risk of harm[16]. Despite the known risks that this technology poses, a current iteration is already in use in over 100 healthcare systems[17,18]. While this rapid rate of implementation may serve as a testament to the urgency and crisis that is physician burnout, it is misaligned with our understanding of this technology's safety. There remains a large gap in understanding how physicians respond when faced with such erroneous AI-generated drafts. This work identified a very low rate in which experienced practicing primary care physicians were able to detect and fix the errors and safety risks in AI-generated draft responses to patient portal messages. In this simulated environment, all but one physician 'sent' at least one fictitious response to a patient that contained an error.

We have identified at least four plausible explanations for why the physicians did not fix the erroneous AI-generated drafts. First, it is possible that providers fell into a functional fixedness[19] trap, a cognitive bias limiting the ability to innovate on an expected solution. As the diagnoses in AI-generated drafts were plausible (e.g., a stomach bug instead of DKA), it may have been more difficult to consider other, less common possibilities. Second, confirmation bias[20] may be at play. This cognitive bias may have led to providers being over-trusting of AI-generated drafts that match their pre-existing expectations of how they would answer the patient message. Third, automation complacency[21], the tendency to monitor automated systems poorly, may have played a role. As the occurrence of automation complacency increases with reliability[22], the fact that most AI-generated drafts did not contain an error or patient safety risk may have resulted in vigilance decrement. Fourth, providers may have succumbed to automation bias[23], the tendency to over-rely on automation, which can be exacerbated under increased workload[24]. Providers simply may have been less diligent in order to reduce their heavy workload. Our finding that providers have positive perceptions of AI-generated drafts (a strong majority agreed that the AI-generated drafts reduced cognitive workload, were accurate, and were safe to use), supports all four explanations as likely contributing to the physician oversight. These results highlight how physicians, as well as other stakeholders, understandably have a growing appetite for AI driven technologies that can address workload burden and may in fact be so overextended that any technology promising a reprieve is eagerly adopted despite the risks. To ensure the safe and effective integration of AI-generated drafts, design improvements, technological advancements, and training interventions will need to be developed. In addition, guidelines for use of these technologies and human review of AI content may need to be adopted at organizational, state, and/or federal levels to safeguard patient care.

There are limitations to the study. Participants were not provided full patient records which may have impacted their ability to appropriately respond to the patient messages as they did not have the additional context they often would in practice. Participants completed the study remotely

Table 2 | Number of misses, number of unedited responses sent, and results of the binomial test for each of the erroneous AI-generated drafts

Error Type	# of Participants who Insufficiently Addressed the Erroneous Message "Misses"	# of Participants who Submitted the AI Draft Unedited	Likelihood that error is missed > 0 P-Value
Objective – Typo	14 (70%)	7 (35%)	<0.001*
Objective – Outdated Advice	14 (70%)	9 (45%)	<0.001*
Potentially Harmful – Blood Clot Risk	13 (65%)	7 (35%)	<0.001*
Potentially Harmful – DKA Risk	15 (75%)	8 (40%)	<0.001*

*significant at adjusted alpha = 0.0125.

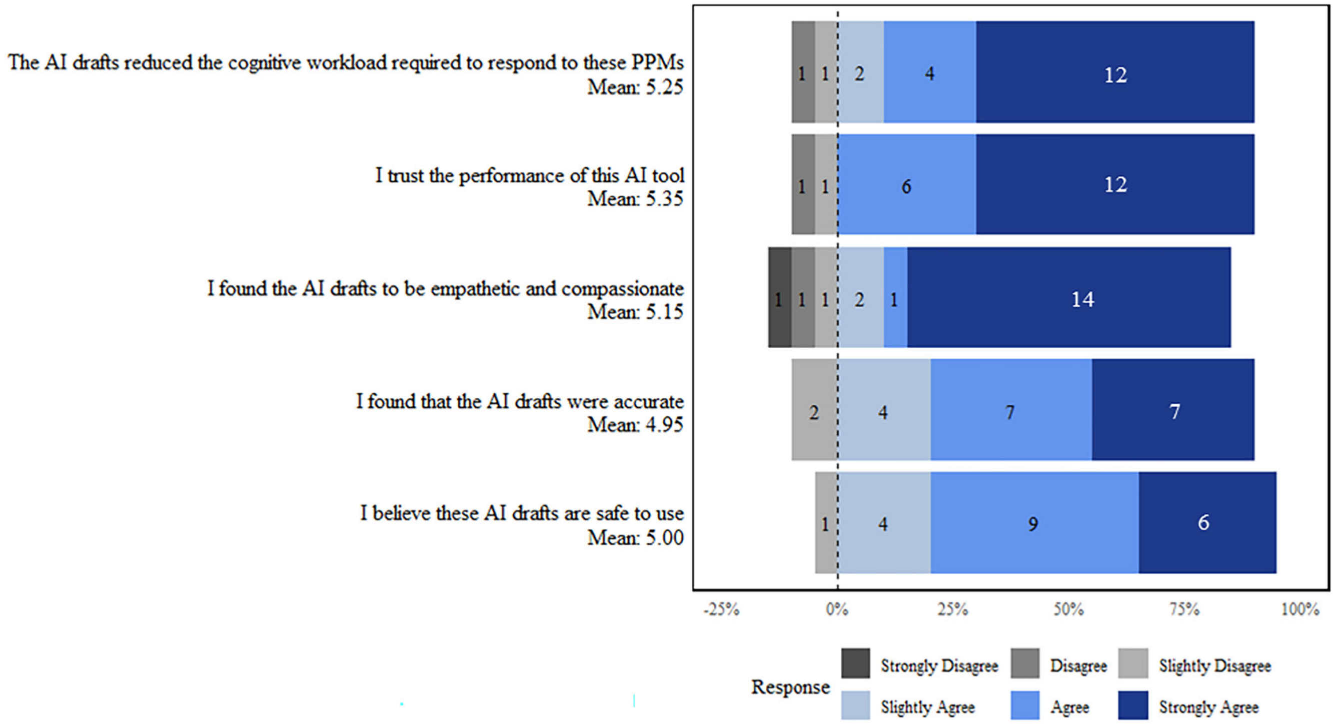

Fig. 1 | Participants agreement with statements regarding the value and safety of using AI drafts to respond to patient portal messages. Mean agreement is calculated by providing a numeric value to the responses (1 = strongly disagree; 6 = strongly agree).

without supervision, and it is unclear whether distractions or other extenuating circumstances may have influenced the results. We did not assess participants' responses to the same PPMs without an AI draft response, so we were not able to compare the error rate that may exist in responses to PPMs without the aid of AI drafts. In practice, physicians encounter a much wider variety of message content than was covered in the 18 PPMs included in this study, future work should examine how results of this work may change for PPMs of varying purpose, content, and complexity. The survey we used was not a psychometrically validated instrument, future work should further investigate similar topics (e.g., reduction in cognitive burden, trust) more thoroughly with validated instruments. Future work should additionally examine how provider characteristics (e.g., age, experience, specialty) correlate with ability to detect and address errors in AI-generated drafts. One of the erroneous questions included a pediatric query, with a mother asking about their child, which the Family Medicine or Pediatric trained doctors may have been more equipped to recognize than the Internal Medicine trained doctors. However, DKA occurs commonly and with the same symptoms for adults, and thus its urgency should be recognized by all of the participants. Additionally, participants who edited the AI response to directly inform the patient to call the child's pediatrician's office was considered to be sufficient.

To better support the use of AI enabled technologies for practice efficiency, additional research is needed to identify the specific types of errors LLMs are likely to make, and the context under which these errors are most prevalent. This would allow for the development of AI systems that may be able to detect errors, interface manipulations that can make salient aspects of messages that may be erroneous, and a host of other potential solutions. Further, specific human factors and cognitive aspects that are responsible for physicians missing errors need to be elucidated. As AI continues to be integrated into clinical practice, cognitive aspects such as functional fixedness, complacency, automation bias, and deskilling will be critical to understand and address.

Methods

This study was approved by the MedStar Health Institutional Review Board (Study ID: 00007728). Informed consent was obtained from all participants. In October 2024, practicing primary care physicians (PCPs) in the Baltimore-Washington metropolitan area were recruited to participate via email. Twenty practicing PCPs from 13 different clinical sites were asked to respond to 18 patient portal messages (PPMs) as they would in their typical clinical role. An AI-generated draft response was provided for each PPM which could be edited directly by the participant. A sample size of 20 participants was pragmatically chosen to balance feasibility with capturing variability in PCP responses.

To create the 18 PPMs, the research team examined 2000 real patient portal messages that patients had sent to primary care physicians. Messages that contained a medical question that could be answered independently of additional patient context that relied on the patient chart were selected to serve as the foundation for the simulated messages. Messages focused on logistical inquiries (e.g., appointment requests, billing questions) were not included in this study. The final 18 simulated PPMs addressed a variety of medical contexts, such as medication side effects, concerning symptoms, chronic conditions, sexually transmitted infections, and more. An AI-generated draft response to each PPM was generated by ChatGPT 4.0 in April 2024. All AI-generated drafts were reviewed by a practicing primary care physician with over 20 years of experience (JW) to identify potential errors. Four (22%) of these PPMs were identified as erroneous. A second practicing primary care physician with over 10 years of experience (AV) reviewed the erroneous AI drafts and provided a consensus opinion. These erroneous messages were then categorized via team consensus: two were identified as objective inaccuracies, and two were identified as potentially harmful omission (Table 3).

Participants responded to the PPMs in a simulated environment which was developed to emulate the aesthetics and functionality of Oracle Cerner's patient messaging system which was familiar to all the participants (Fig. 2). The order in which participants responded to the messages was randomized to control for learning effects. The simulated environment captured all participant changes to each AI-generated draft response and the time spent answering each message. Participants' edits to the erroneous AI-generated draft responses and the final submitted responses were independently evaluated by two practicing primary care physicians (JW, AV) to identify whether

Table 3 | Errors in the AI-generated draft responses to patient portal messages

Error Type	Description of Error
Objective – Typo	AI draft incorrectly spells a medication (which the patient themselves misspelled). E.g., "Thank you for reaching out about adjusting your Centrizine prescription."
Objective – Outdated Advice	AI draft provides advice regarding the timing of getting a COVID vaccination after testing positive which differs from current CDC guidance. E.g., "As for getting vaccinated after having COVID-19, the current guidance generally suggests that individuals can be vaccinated as soon as they are out of isolation and their symptoms have resolved." Current CDC guidance states that one may choose to delay receiving a COVID-19 booster until 3 months after a recent COVID-19 infection [25].
Potentially harmful – Blood Clot Risk	Patient mentions edema and red splotches at the ankle. While the AI draft does mention the risk of blood clot and to make an appointment as soon as possible, it does not urgently inform the patient to seek an evaluation at urgent or emergency care to address the risk of a blood clot.
Potentially harmful – DKA Risk	Patient mentions that their child has been vomiting, drinking a lot of water, and "still hasn't figured out her bedtime problems" suggesting urination issues. The AI draft suggests a possible stomach virus or infection, and suggests bringing the child in for a consultation if vomiting continues. The AI draft does not address the risk of diabetic ketoacidosis which requires urgent intervention.

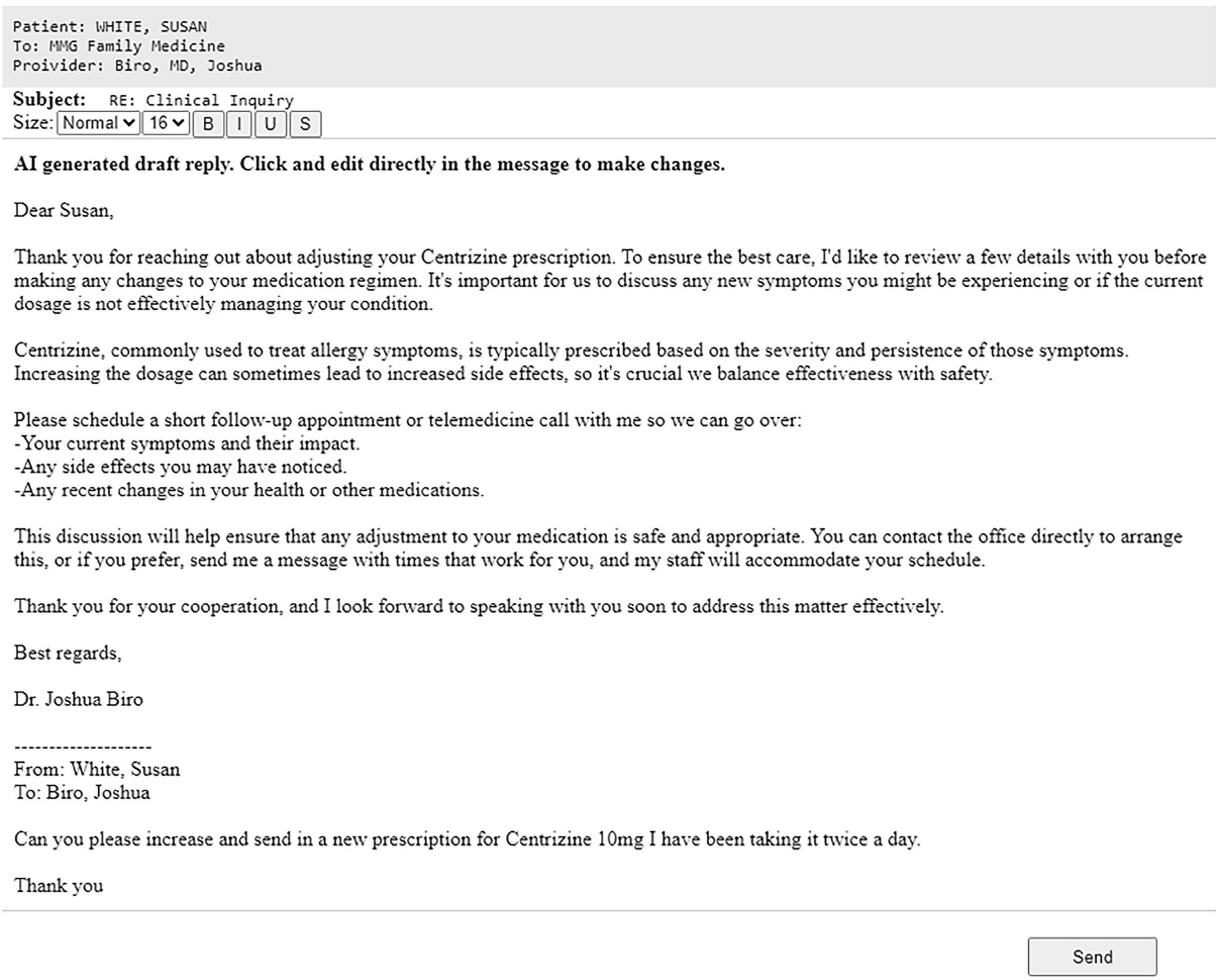

Fig. 2 | Screenshot of the simulated environment once a message was selected. The AI-generated draft can be edited directly prior to the message being sent back to the patient.

the responses were edited to address the potential harm. Inter-rater agreement was 92.5% with substantial agreement indicated by Cohen's Kappa ($\kappa = 0.81$). Only responses which were identified as 'insufficiently edited to address the error' by both primary care physicians were labeled as a "miss".

After responding to all 18 PPMs, participants were administered a survey asking their opinions on the value and safety of the AI-generated PPM response drafts. The survey included one 'True or False' question: *"I found the AI drafts to be helpful in responding to these PPMs,"* and five 6-point Likert-scale questions asking participants to rate their level of agreement (1 representing strongly disagree, 2 disagree, 3 slightly disagree, 4 slightly agree, 5 agree, and 6 strongly agree) with the following statements:

1. *"The AI drafts reduced the cognitive workload required to respond to these PPMs."*
2. *"I found the AI drafts to be empathetic and compassionate."*

3. *"I trust the performance of this AI tool."*
4. *"I found that the AI drafts were accurate."*
5. *"I believe these AI drafts are safe to use."*

Descriptive statistics on the number of misses for each question and the responses to the Likert-scale questions are provided. Our primary outcome was the likelihood of a missed error (i.e., whether an error in the AI-generated draft response remains present after physician review). To examine if the likelihood of a miss is significantly greater than zero, we conducted a separate binomial test for each of the four erroneous AI-generated drafts. To control for Type 1 error across multiple comparisons, a Bonferroni correction was applied resulting in an adjusted significance level of alpha = 0.0125.

Data availability
The datasets generated and analyzed as part of this study will be made available by the corresponding author upon reasonable request.

Code availability
The code used to analyze the datasets are available from the corresponding author upon reasonable request. Statistics were performed using the open-source statistical software R.

Received: 17 December 2024; Accepted: 26 March 2025;
Published online: 24 April 2025

References
1. North, F. et al. A retrospective analysis of provider-to-patient secure messages: How much are they increasing, who is doing the work, and is the work happening after hours? *JMIR Med. Inf.* **8**, e16521 (2020).
2. Nath, B. et al. Trends in electronic health record inbox messaging during the COVID-19 pandemic in an ambulatory practice network in New England. *JAMA Netw. Open* **4**, e2131490 (2021).
3. Holmgren, A. J. et al. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. *J. Am. Med. Inform. Assoc.* **29**, 453–460 (2022).
4. Martinez, K. A., Schulte, R., Rothberg, M. B., Tang, M. C. & Pfoh, E. R. Patient portal message volume and time spent on the EHR: an observational study of primary care clinicians. *J. Gen. Intern. Med.* **39**, 566–572 (2024).
5. Arndt, B. G. et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann. Fam. Med.* **15**, 419–426 (2017).
6. Yan, Q., Jiang, Z., Harbin, Z., Tolbert, P. H. & Davies, M. G. Exploring the relationship between electronic health records and provider burnout: a systematic review. *J. Am. Med. Inform. Assoc.* **28**, 1009–1021 (2021).
7. Budd, J. Burnout related to electronic health record use in primary care. *J. Prim. Care Community Health* **14**, 215013192311669 (2023).
8. Adler-Milstein, J., Zhao, W., Willard-Grace, R., Knox, M. & Grumbach, K. Electronic health records and burnout: time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. *J. Am. Med. Inform. Assoc.* **27**, 531–538 (2020).
9. Baxter, S. L. et al. Association of electronic health record inbasket message characteristics with physician burnout. *JAMA Netw. Open* **5**, e2244363 (2022).
10. Shanafelt, T. D. 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.* **97**, 491–506 (2022).
11. Lieu, T. A. et al. Primary care physicians' experiences with and strategies for managing electronic messages. *JAMA Netw. Open* **2**, e1918287 (2019).
12. Murphy, D. R., Satterly, T., Giardina, T. D., Sittig, D. F. & Singh, H. Practicing clinicians' recommendations to reduce burden from the electronic health record inbox: a mixed-methods study. *J. Gen. Intern. Med.* **34**, 1825–1832 (2019).
13. Tai-Seale, M. et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open* **7**, e246565 (2024).
14. Garcia, P. et al. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Netw. Open* **7**, e243201 (2024).
15. Yan, S. et al. Prompt engineering on leveraging large language models in generating response to InBasket messages. *J. Am. Med. Inform. Assoc.* **31**, 2263–2270 (2024).
16. Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health* **6**, e379–e381 (2024).
17. Landi, H. Epic touts new software features 'turbocharged with AI'. https://www.fiercehealthcare.com/ai-and-machine-learning/epic-touts-new-ai-applications-streamline-charting-and-bring-research (2024).
18. Gen A. I. Saves nurses time by drafting responses to patient messages. *EpicShare.* https://www.epicshare.org/share-and-learn/mayo-ai-message-responses (2024).
19. Adamson, R. E. Functional fixedness as related to problem solving: a repetition of three experiments. *J. Exp. Psychol.* **44**, 288–291 (1952).
20. Klayman, J. Varieties of confirmation bias. in *Psychology of Learning and Motivation*, Vol. 32 (eds. Busemeyer, J., Hastie, R. & Medin, D. L.) 385–418 (Academic Press, 1995).
21. Parasuraman, R., Molloy, R. & Singh, I. L. Performance consequences of automation-induced 'complacency'. *Int. J. Aviat. Psychol.* **3**, 1–23 (1993).
22. Molloy, R. & Parasuraman, R. Monitoring an automated system for a single failure: vigilance and task complexity effects. *Hum. Factors* **38**, 311–322 (1996).
23. Skitka, L. J., Mosier, K. L. & Burdick, M. Does automation bias decision-making? *Int. J. Hum. Comput. Stud.* **51**, 991–1006 (1999).
24. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med Inf. Assoc.* **19**, 121–127 (2012).
25. CDC. Staying up to date with COVID-19 Vaccines. *COVID-19*. https://www.cdc.gov/covid/vaccines/stay-up-to-date.html (2024).

Acknowledgements
This research was funded by the Agency for Healthcare Research and Quality (Award #: 1R03HS030307-01).

Author contributions
J.B. developed the study design, interpreted and analyzed the data, and was the primary author for the manuscript. J.H. contributed to study design, data interpretation, drafting, and critically reviewing the manuscript. M.M. was the primary designer of the simulated environment for testing, in addition to contributing to study design and critically reviewing the manuscript. A.V. and J.W. provided clinical perspectives to the study design and were the judges for determining whether the PPM responses were sufficient, in addition to critically reviewing the manuscript. R.R. and G.T. contributed to study conception, study design, data interpretation, and critically reviewing the manuscript.

Competing interests
Author R.R. serves as an editor of this journal but had no role in the peer-reviewed decision to publish this manuscript. All other authors declare no financial or non-financial competing interests.

Additional information
Correspondence and requests for materials should be addressed to Joshua M. Biro.