

Automating Responses to Patient Portal Messages Using Generative AI

Amarpreet Kaur, MHS¹, Alexander Budko (COL 2026)², Katrina Liu (COL 2025)², Eric Eaton, PhD², Bryan Steitz, PhD³, Kevin B. Johnson, MD, MS^{1,2}

Funded by: Penn Undergraduate Research Mentoring Program

1 Perelman School of Medicine, University of Pennsylvania, 2 School of Engineering and Applied Science, University of Pennsylvania, 3 Vanderbilt University Medical Center, Vanderbilt University



INTRODUCTION

- Patient portals are vital in modern healthcare, facilitating secure access to health information and communication between healthcare professionals and patients.

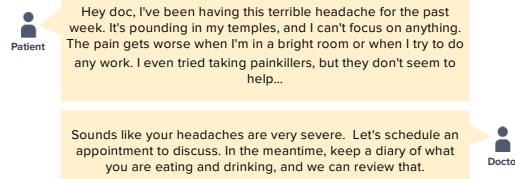


Fig 1. Diagram representation of a patient-doctor dialogue using the EHR system's in-basket messaging feature (not based on actual data).

- The rapid adoption of patient portal messaging, especially during the COVID-19 pandemic, has led to a surge in in-basket patient messages, exacerbating healthcare provider burnout.¹⁻⁵
- Generative AI, like OpenAI's GPT-4, offers a promising solution to alleviate documentation-related burnout by automating tasks such as generating replies to patient in-basket inquiries, drafting clinic letters and assisting in medical notes.⁷⁻¹⁰
- This study aims to explore healthcare provider acceptability of AI-generated message responses, offering insights into the feasibility of integrating large language models (LLMs) to enhance efficiency in patient communication and alleviate workload burdens on clinicians.

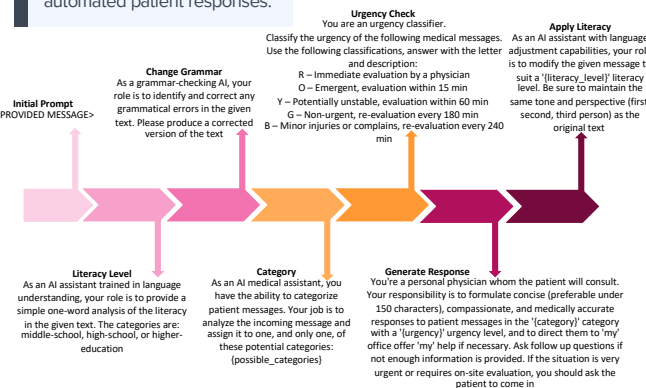
METHODOLOGY

Initial Synthetic Patient-Message Data Collection

1. Retrieved 85 patient portal messages and clinician responses from Vanderbilt University Medical Center (VUMC) repository.
2. De-identified and manually rephrased messages to maintain content while varying tone and length.
3. Engineered a prompt within GPT-4 to generate messages similar in tone, length, and topic to the originals.
4. Recruited eight clinicians to review and compare synthetic and authentic patient portal messages, finding that clinicians correctly identified clinician-generated messages only 51.1% of the time.
5. Combined the message pool to develop synthetic patient portal message responses based on the study's findings.

Pipeline Development

The diagram below depicts the engineering of the pipeline responsible for generating automated patient responses.



RESULTS

Response Assessment: Real vs. GPT

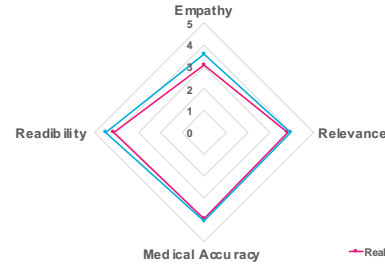


Figure 2: Evolution of the Pipeline. The radar diagram illustrates the mean comparison of GPT-generated and real responses using an ordinal scale ranging from 1 (low) to 5 (high). A rating of 1 indicates poor performance, while 5 signifies excellent performance.

	GPT Response		Real Response		Significance (t-test)
	Mean (+/- SD)	Median	Mean (+/- SD)	Median	
Empathy	3.57 (1.02)	3.6	3.07 (1.00)	3.1	< 0.001
Relevance	3.94 (1.00)	4.2	3.81 (1.09)	4	0.08
Medical Accuracy	4.05 (0.92)	4.2	3.95 (0.99)	4	0.12
Readability	4.50 (0.68)	4.9	4.13 (1.01)	4.7	< 0.001

Table 1: Comparative Analysis of GPT versus real message responses. The table above provides a comprehensive breakdown of the average means and medians derived for the four key characteristics, comparing GPT-generated message-response pairs to real ones. Both empathy and readability were statistically better for GPT-generated responses.

MESSAGE RESPONSE PAIR EXAMPLES

Patient: Can you tell me what type of test this is: Natriuretic Peptide, as it appears that I am well over the range. Thanks.

GPT-4: The Natriuretic Peptide test is used to measure the levels of a hormone that is released by the heart when it is under stress. If the levels are higher than normal, it could be a sign of heart failure or other heart-related conditions. We kindly request that you schedule an appointment to discuss your test results in more detail.

Example of a GPT-4 generated response

Patient: This thing in my throat is continuing to progress. I now have a steady dull pain along with a few random stabbing pains in that area. Swallowing has become a bigger issue also. I have an appointment with you this Friday, and with speech pathology afterwards. I am ready to get this thing out.

Doctor: I tried to call you several times today. I am sorry to have not gotten in touch. I hope that your symptoms do not worsen rapidly. If they do, you can come to ER here at anytime day or night and we will take care of you.

Example of a response written by a real provider

CONCLUSION

The findings of this study suggest that GPT-4 generated responses are feasible and acceptable to primary care providers. Despite the small sample size and single healthcare system representation, the study provides promising insights into the potential of AI-driven messaging systems to alleviate clinician burnout and enhance patient communication. As with all technological endeavors, continual evolution is paramount for addressing challenges and leveraging emerging insights from both the technological and cognitive domain.

Evaluation of Message Response Pairs

The evaluation of GPT-4 generated patient responses was conducted through a 20-question survey sent to 49 health care workers across the University of Pennsylvania Health System. Survey assessed message quality and authenticity across four dimensions:

1. **Empathy** reflecting the degree of consideration for the patient's emotions in the message
2. **Relevance** assessing how closely the content addressed the patient's expressed needs
3. **Medical Accuracy** gauging the alignment of the message with established medical practices and guidelines
4. **Readability** evaluating the clarity, coherence, and simplicity of the language employed

Participants were also tasked to identify if message response were written by GPT or a human provider, gauging authenticity human-like nature of the generated responses and evaluating GPT-4's medical interaction quality.

REFERENCES

1. Carré E, Vilari L, Pizzullo AM, et al. The Impact of Digital Patient Portals on Health Outcomes, System Efficiency, and Patient Attributes: Updated Systematic Literature Review. *J Med Internet Res*. 2023;25(2):e36889. doi:10.2196/36889
2. Hinchings AJ, Downing NL, Tang M, Sheep C, Longhurst C, Haslam RC. Assessing the Impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. *Journal of the American Medical Informatics Association*. 2022;29(3):453-460. doi:10.1093/jamitab/39
3. Tai-Seale M, Baxter S, Miller M, et al. Association of physician burnout with perceived EHR work stress and potentially extensible factors. *Journal of the American Medical Informatics Association*. 2022;29(9):965-972. doi:10.1093/jamitab/38
4. Johnson KB, Neuss MJ, Dettmer DE. Electronic health records and clinician burnout: A study of three sites. *J Am Med Inform Assoc*. 2021;28(5):967-973. doi:10.1093/jamia/ocaa274
5. Johnson KB, Ibrahim SA, Rosenbloom ST. Enabling Equitable Access to Patient Portals—Closing the Technology Gap. *JAMA Health Forum*. 2023;4(7):e233406
6. Liu S, McCoy AB, Wright AP, et al. Leveraging Large Language Models for Generating Responses to Patient Messages. *Health Informatics*. 2023. doi:10.1101/2023.07.14.23292649
7. Liu S, Wright AP, Peterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*. 2023;30(1):27-34. doi:10.1093/jamia/ocac072
8. Rajulu R, Aranyne JS, Zaidat B, et al. ChatGPT and its Role in the Decision-Making for the Diagnosis and Treatment of Lumbar Spinal Stenosis: A Comparative Analysis and Narrative Review. *Global Spine Journal*. Published online August 15.
9. 2023;29(5):622-629. doi:10.1177/10939862231195763
10. Kao HJ, Chen TH, Wang WC, Chen W, Chen JC. Assessing ChatGPT's capacity for clinical decision support in pediatrics: A comparative study with pediatricians using NIDMAP of Rasch analysis. *Medicine*. 2023. doi:10.1093/med/ckad036
11. Ayres AJ, Pissak A, Drexler 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. doi:10.1001/jamainternmed.2023.1838