

Applied Clinical Informatics

Primary Care Providers Acceptance of Generative AI Responses to Patient Portal Messages

Amarpreet Kaur, Alex Budko, Katrina Liu, Bryan D. Steitz, Kevin B Johnson.

Affiliations below.

DOI: 10.1055/a-2565-9155

Please cite this article as: Kaur A, Budko A, Liu K et al. Primary Care Providers Acceptance of Generative AI Responses to Patient Portal Messages. ACI 2025. doi: 10.1055/a-2565-9155

Conflict of Interest: The authors declare that they have no conflict of interest.

This study was supported by National Institutes of Health (<http://dx.doi.org/10.13039/100000002>), 5DP1LM014558

Abstract:

Background: Patient portals bridge patient and provider communications but exacerbate physician and nursing burnout. Large language models (LLMs) can generate message responses that are viewed favorably by healthcare professionals; however, these studies have not included diverse message types or new prompt-engineering strategies. Our goal is to investigate and compare the quality and precision GPT-generated message responses versus real doctor responses across the spectrum of message types within a patient portal.

Methods: We used prompt engineering techniques to craft synthetic provider responses tailored to adult primary care patients. We enrolled a sample of primary care providers in a cross-sectional study to compare authentic with synthetic patient portal message responses, generated by GPT-3.5-turbo, July 2023 version (GPT). The survey assessed each response's empathy, relevance, medical accuracy, and readability on a scale from 0 to 5. Respondents were asked to identify responses that were GPT-generated vs. provider-generated. Mean scores for all metrics were computed for subsequent analysis.

Results: A total of 49 health care providers participated in the survey (59% completion rate), comprising 16 physicians and 32 advanced practice providers (APPs). In comparison to responses generated by real doctors, GPT-generated responses scored statistically significantly higher than doctors in two of the four parameters: empathy ($p < 0.05$) and readability ($p < 0.05$). However, no statistically significant difference was observed for relevance and accuracy ($p > 0.05$). Although readability scores were significantly different, the absolute difference was small, and the clinical significance of this finding remains uncertain.

Conclusion: Our findings affirm the potential of GPT-generated message responses to achieve comparable levels of empathy, relevance, and readability to those found in typical responses crafted by healthcare providers. Additional studies should be done within provider workflows and with careful evaluation of patient attitudes and concerns related to the ethics as well as the quality of generated responses in all settings.

Corresponding Author:

Dr. Kevin B Johnson, University of Pennsylvania, Biostatistics, Epidemiology, and Informatics, 3700 Hamilton Walk, 19087 Philadelphia, United States, Kevin.johnson1@penmedicine.upenn.edu

Affiliations:

Amarpreet Kaur, University of Pennsylvania, Biostatistics, Epidemiology, and Informatics, Philadelphia, United States

Alex Budko, University of Pennsylvania, Computer and Information Sciences, Philadelphia, United States

Katrina Liu, University of Pennsylvania, Computer and Information Science, Philadelphia, United States

Bryan D. Steitz, Vanderbilt University Medical Center, Biomedical Informatics, Nashville, United States



This article is protected by copyright. All rights reserved.

Accepted Manuscript

Automating Responses to Patient Portal Messages Using Generative AI

Amarpreet Kaur, MHS

Clinical Research Coordinator, Department of Biostatistics, Epidemiology, and Informatics

Perelman School of Medicine, University of Pennsylvania

B202 Richards Building, 3700 Hamilton Walk

University of Pennsylvania, Philadelphia, PA 19104-6116

amarpreet.kaur@penmedicine.upenn.edu

Alexander Budko

School of Engineering and Applied Science, University of Pennsylvania

202 Augusta Way

Melbourne, FL 32940

abudko@seas.upenn.edu

Katrina Liu

School of Engineering and Applied Science, University of Pennsylvania

B202 Richards Building, 3700 Hamilton Walk

University of Pennsylvania, Philadelphia, PA 19104-6116

katltn@seas.upenn.edu

Eric Eaton, PhD

School of Engineering and Applied Science, University of Pennsylvania

Levine Hall, Levine 464

3330 Walnut Street

Philadelphia, PA 19104-6309

eeaton@seas.upenn.edu

Bryan D. Steitz, PhD

Vanderbilt University Medical Center, Vanderbilt University

2525 West End, Suite 1475

Nashville, TN 37203

bryan.d.steitz@vumc.org

Kevin B. Johnson, MD, MS, *corresponding author*

Perelman School of Medicine, University of Pennsylvania

B202 Richards Building, 3700 Hamilton Walk

University of Pennsylvania, Philadelphia, PA 19104-6116

kevin.johnson1@penmedicine.upenn.edu

ABSTRACT

Background: Patient portals bridge patient and provider communications but exacerbate physician and nursing burnout. Large language models (LLMs) can generate message responses that are viewed favorably by healthcare professionals; however, these studies have not included diverse message types or new prompt-engineering strategies. Our goal is to investigate and compare the quality and precision GPT-generated message responses versus real doctor responses across the spectrum of message types within a patient portal.

Methods: We used prompt engineering techniques to craft synthetic provider responses tailored to adult primary care patients. We enrolled a sample of primary care providers in a cross-sectional study to compare authentic with synthetic patient portal message responses, generated by GPT-3.5-turbo, July 2023 version (GPT). The survey assessed each response's empathy, relevance, medical accuracy, and readability on a scale from 0 to 5. Respondents were asked to identify responses that were GPT-generated vs. provider-generated. Mean scores for all metrics were computed for subsequent analysis.

Results: A total of 49 health care providers participated in the survey (59% completion rate), comprising 16 physicians and 32 advanced practice providers (APPs). In comparison to responses generated by real doctors, GPT-generated responses scored statistically significantly higher than doctors in two of the four parameters: empathy ($p < 0.05$) and readability ($p < 0.05$). However, no statistically significant difference was observed for relevance and accuracy ($p > 0.05$). Although readability scores were

significantly different, the absolute difference was small, and the clinical significance of this finding remains uncertain.

Conclusion: Our findings affirm the potential of GPT-generated message responses to achieve comparable levels of empathy, relevance, and readability to those found in typical responses crafted by health care providers. Additional studies should be done within provider workflows and with careful evaluation of patient attitudes and concerns related to the ethics as well as the quality of generated responses in all settings.

Keywords:

1. Patient Web Portal
2. Physician-Patient Interaction
3. Artificial Intelligence
4. Communication, Health
5. Medical Informatics
6. Large language models

INTRODUCTION

Patient portals have become an integral and indispensable component of modern healthcare, providing patients with secure online access to vital health information and facilitating crucial communication bridges between healthcare professionals (HCPs) and patients. In doing so, they foster stronger connections between providers and patients and facilitate the delivery of personalized care through effective communication.¹

However, with the rapid growth of patient portals, especially during the COVID-19

pandemic (2), this convenience has led to an overwhelming surge in the number of in-basket messages healthcare providers must manage daily. This growing volume of these in-basket messages, many of which are administrative or routine, contribute to clinical burnout, reduce efficiency, and divert valuable time from direct patient care.¹⁻⁴ First documented in 1974, physician burnout has been linked to the demands of EHR documentation, consuming substantial clinical time.^{2,5} Primary care providers face uniquely heightened burnout risks among all HCPs, emphasizing the pressing need for interventions to alleviate EHR-related burdens and support clinician well-being.²

Generative AI tools (GenAI) such as OpenAI®'s GPT-4, have emerged as promising tools in the healthcare sector, particularly for mitigating documentation-related burnout among clinicians. GenAI has gained widespread attention in the medical community due to its capacity to effectively generate patient clinic letters, radiology reports, medical notes, discharge summaries, and clinical decision support.⁶⁻¹⁰ Researchers have begun to explore the patient-facing role of GenAI in areas such as chatbots^{11,12}, wearable technologies¹³⁻¹⁵ and automating patient portal message responses^{9,16}. In particular, a recent study by Tai-Seale with 52 participants showed that GenAI drafts took longer to read, were about 18% longer in length, and while responses were less detailed, they were generally more empathetic. Most respondents edited messages to remove recommendations for appointments or inaccuracies. This study did not divide messages into message types, such as those described by Sulieman and colleagues¹⁷, thereby not addressing if specific types of messages may require less scrutiny or human input. It is critical to note that while, GenAI has shown promise in automating routine or administrative messages and easing the load on healthcare providers, not all patient communications are suitable for automation. Understanding which types of patient

messages may require less scrutiny or human input is critical to ensuring both efficiency and safety in healthcare delivery.

Currently, there is limited research and guidance on how to triage patient messages for potential automation, creating an essential gap in knowledge that can affect the effective integration of AI tools. This study aims to address this gap by evaluating the appropriateness of automating responses to specific message types while also exploring healthcare provider acceptability of GenAI in daily clinical practices. By identifying which messages can be safely handled by AI versus those requiring human attention, this research seeks to optimize healthcare delivery by streamlining communication workflows. The ultimate goal is to alleviate clinician workload, improve response times for patients and ensuring high-quality personalized care remains a priority, thus offering a practical solution to a growing issue in modern healthcare.

METHODS

Study Setting

This cross-sectional, single-group study was conducted at a major urban integrated academic medical center in the Northeast. We primarily focused on physicians and advanced practice providers in internal medicine and family medicine. Healthcare providers in this setting are located in over 50 locations and function as a patient-centered medical home for millions of patients. The study was deemed nonhuman participants research by the University of Pennsylvania Institutional Review Board. Answering the survey served as implied consent to participate in the project.

Initial Data Collection and Creation of Synthetic Patient Portal Messages

Considering the sensitive nature of real patient portal messages, we first retrieved a set of 85 patient portal message-response pairs from a repository at Vanderbilt University Medical Center. This set of 85 contained messages from both the provider and the patient, resulting in 170 total messages. All identifying language was removed from each message, with identifying nouns replaced with semantically similar language, allowing us to use them on commercial cloud platforms. No changes were made in tone, urgency or vernacular. Two investigators (BDS, KBJ) then independently categorized each patient message into the categories as described by Sulieman¹⁷ (management, interventions, problems, referrals, test results, clinical intervention preparation). Interrater agreement for the categorization of these messages was nearly perfect (Kappa = 0.94).

Using the authentic provider responses to these rephrased patient messages, we engineered a prompt within GPT-3.5-turbo, July 2023 version (GPT)¹⁸ to generate additional message responses similar in tone, length, and word choice. We used a strategy based on “chain-of-thought” reasoning, which facilitated the incorporation of explanatory iterative changes in text generated by large language models.^{19,20} In our use case, we used this strategy to tailor an initial GPT response to the tone and language of the original message. For example, we can modify the response length of a message without compromising accuracy by iteratively modifying a component of the prompt pipeline, as shown in Figure 1. A convenience sample of eight experienced primary care physicians reviewed these GPT-generated messages and were unable to distinguish them from the original patient messages.

Given these results, we combined our pool of messages into one set to develop our synthetic patient portal message responses for the second phase of the study.

Pipeline Development

We used iterative prompt engineering, similar to chain of thought, to create tailored provider responses to patient messages. These engineering prompts contained explicit instructions to mimic the tone, word choices (slang instead of medical words) and brevity of each rephrased patient message but did not include any content from the patient messages used in the test set for GPT-generated responses. Once we were satisfied with the face validity of responses, we generated synthetic patient portal message responses across the range of categories described above, using this prompt:

“You're a health care provider whom the patient will consult. Your responsibility is to formulate concise (preferably under 150 characters), compassionate, and medically accurate responses to patient messages in the '{category}' category with a '{urgency}' urgency level, and to direct them to 'my' office offer 'my' help if necessary. Ask follow up questions if not enough information is provided. If the situation is very urgent or requires on-site evaluation, you should ask the patient to come in.”

The final pipeline is summarized in Figure 1 and all code to generate the messages is available on Github.²¹

Evaluation of Message Response Pairs

To evaluate the quality and authenticity of messages generated by our pipeline, we conducted a cross-sectional survey study of healthcare providers (HCPs) across the

University of Pennsylvania. The survey consisted of 20 questions, each including a message-response pair. Participants were presented with a total of 20 unique patient messages, with each message accompanied by a provider response. Of these 20 unique patient message-provider response pairs, Ten responses were generated de novo using GPT, while the remaining ten were written by real doctors. Participants were not informed which responses were generated by GPT and which were provided by real doctors, ensuring that their evaluations were based solely on the quality of the responses, independent of their source. For each question, participants were asked to rate the message-response pairs according to four key quality dimensions of communication, shown by Ayers and colleagues.²² These quality dimensions aimed to predict the perceived quality of the response: *Empathy*, reflecting the degree of consideration for the patient's emotions in the response; *Relevance*, assessing how closely the response addressed the patient's expressed needs; *Medical Accuracy*, gauging the alignment of the response with established medical practices and guidelines; and *Readability*, evaluating the clarity, coherence, and simplicity of the language employed. Each quality dimension was presented as a Likert-scale question with five possible responses, ranging from 0 to 5. Additionally, participants were asked to discern whether each message response was GPT-generated or written by a real provider.

We recruited HCPs who identified as primary care MDs, DOs, or advanced practice providers (APPs), through an email distribution list to complete the survey. This sampling frame covered most primary care providers at our institution. Eighty-four potential participants responded to the initial email request and a survey link was sent to each interested participant. Upon completion of the survey, participants received a \$10

Starbucks gift card as a token of appreciation. The survey was distributed using both Google Forms and REDCap, the latter being utilized due to firewall restrictions preventing Google Forms use on some computers. The survey was administered between November 28, 2023, and January 5, 2024. We used JMP (version 17.2.0)¹⁷ to complete univariate analyses and to determine if participants who could discern which messages were generated by GPT differed in their assessment of message quality.

RESULTS

Of the 84 HCPs who showed interest in participating in the survey, 49 completed the survey, resulting in a 58% completion rate. Table 1 provides an overview of various demographic and professional variables among the 49 respondents. Most participants identified as female (77.6%), with 69% between the ages of 31 and 40. A total of 67% of respondents identified as APPs, while 33% held a medical degree (MD or DO). Years of experience seeing patients varied, with the largest group having less than five years of experience (31%), followed by experience between 10-15 years (18%). Most respondents worked in clinics (69%), in urban settings (63%), and reported receiving 25-75 in-basket messages from patients during a typical work week (55%). Most respondents (76%) indicated no or unknown experience with AI tools in medical practice.

Table 2 and figure 2 summarize the overall assessment of message-response quality. Notably, GPT-generated responses generally outperformed real responses across all key characteristics, demonstrating statistical significance with empathy ($p < 0.001$) and readability ($p < 0.001$). Relevance also trended toward significance ($p = 0.08$).

Participants correctly identified GPT messages 73% of the time (good guessers) and

correctly identified authentic messages 50% of the time. There were no statistically significant differences in assessed message-response quality for good guessers versus other participants as determined by one-way ANOVA ($F(1,47) = 2.27, p = 0.13$).

DISCUSSION

In this study, HCPs evaluated the quality of synthetic (GPT-generated) versus authentic (provider generated) message responses. The results revealed that responses generated by GPT achieved statistically significant ratings in empathy and readability, with a notable trend toward statistical differences in relevance and medical accuracy compared to authentic patient portal message responses. The AI's capacity to generate empathetic and easily readable messages may stem from the technology's ability to invest unlimited time in crafting responses, as opposed to real doctors, who often face time constraints and stress when generating responses to patient messages. These findings not only build upon but also validate previous research by Ayers and colleagues,⁴ where a small team of healthcare professionals rated online chatbot responses as more empathetic than verified physician responses. Through a more tailored approach to prompt engineering, these results complement those of a recent study by Tai-Seale and colleagues, where participants found the technology favorable, but suggested less “robotic” responses that required less editing.¹⁶ However, it is important to note that our study did not find significant differences in two critical measures—relevance and accuracy—between AI and provider-generated responses. This indicates that while AI may enhance certain aspects of communication, such as tone and readability, it may not yet be superior to human input in areas requiring nuanced judgment, such as the relevance of the response to a patient's medical issue

or the accuracy of clinical information. Moreover, while readability differences were statistically significant, the absolute difference was relatively small, raising questions about the clinical significance of these results. This emphasizes the need for careful interpretation of these findings, particularly when considering the practical impact on patient care.

Our study extends these findings by including more response types and utilizing chain-of-thought reasoning, to create more nuanced and explainable output. Another study by Garcia and colleagues²³ assessed AI-generated reply utilization and found an overall utilization rate of 20% across 162 clinicians in primary and specialty care, despite draft replies being available for more than significantly more messages. Utilization was affected by technical limitations or internal exclusion criteria, preventing the generation of a message response. Themes affecting adoption included tone, content relevance and accuracy. Based on our findings, although the complexity of the prompt engineering to create these messages may be high, that cost may result in improved adoption and is likely to be easier to implement over time.

Additionally, it is critical to recognize that this study was not designed as a non-inferiority trial, meaning we cannot definitively conclude that AI-generated responses are comparable to human responses across all quality parameters. Further research is needed to evaluate whether AI tools can consistently maintain high standards of clinical relevance and safety, especially in more complex patient interactions.

As AI-enabled messaging systems continue to mature and advance, with attention to message tailoring and the specific needs of patients from diverse backgrounds, chatbots and similar tools are likely to become more commonplace in medicine.

Already, several studies are exploring the feasibility of integrating systems such as GPT to generate high-quality responses to patient inquiries and aid clinical decisions-making across various medical specialties.^{7,24,25} While AI has shown promise in alleviating the communication burden on healthcare providers, its use should be considered as a supportive tool rather than a replacement for human expertise. AI may be best suited to handle certain types of patient messages that require less medical nuance but should not yet be fully relied upon for more complex decision-making without human oversight. Of note, crafting effective prompts in our study entailed iterative trial and error. The potential for performance variation underscores the importance of understanding the model's reliance on training data patterns and ensuring the relevance and quality of examples provided. Our resulting strategy and prompts are available for reference, providing valuable insights for future research and implementation endeavors in this rapidly evolving field.

LIMITATIONS

This study is subject to several limitations that may impact its generalizability. Firstly, the sample size of both generated messages (8) and participating survey respondents (49) is small, potentially limiting the breadth of perspectives represented. All participants were drawn from a single healthcare system, which may not fully capture the diversity of opinions regarding the value proposition for patient portal message responses or the preferred format and comprehensiveness of these responses across different healthcare settings. Furthermore, the study relied on a convenience sample of providers who may have had more time and interest to participate in the survey, introducing a

potential bias in the results. As such, caution should be exercised when generalizing the findings of this study to broader populations.

The patient portal messages used to generate these synthetic provider responses were deidentified, with identifying nouns replaced. At the time of this study, we were not permitted to use HIPAA safe harbor compliant messages outside of the health system firewall. We anticipate that health systems will relax this constraint shortly, which will facilitate larger studies within a health system. Finally, our use of prompt engineering to generate responses is currently a trial-and-error process, with features of messages proposed by our research team. Our study was conducted using GPT 3.5-turbo, which is an older and less functional version of GPT than would be used in any clinical trials of this innovation. Our study also tested only the capabilities of GenAI in isolation. Clinical trials of this technology should combine chart content, message threads, and patient preferences, possibly leveraging retrieval-augmented generation or expanding context windows to improve personalization of responses.

It will be important to better understand the desirable characteristics of patient portal message responses from the perspective of health care providers and patients.

Potential Biases

The survey was administered to healthcare providers (HCPs) via Google Forms and REDCap. However, no specific measures were put in place to prevent participants from collaborating or discussing the survey or their responses with one another. As a result, there is a potential for response bias or social desirability bias to have been introduced. HCPs may have been influenced by side conversations or peer expectations, leading to

responses that were shaped by group discussions rather than independent, unbiased opinions, which could affect the validity of the survey results.

Furthermore, while the survey aimed to capture a broad perspective from healthcare providers (HCPs), we did not employ a purposive sampling approach to specifically target participants with varying levels of experience with AI tools in medical practice. This limitation may have introduced sampling bias and influenced the diversity of responses, as HCPs with more experience using AI might have responded differently compared to those with less or no experience. Future studies could benefit from purposive sampling to ensure a more representative sample in terms of AI familiarity, which may provide deeper insights into the technology's impact on clinical practice.

FUTURE WORK

Considering the limitations of our pipeline, several areas for future research and improvement emerge. Quantitative assessments are crucial to validate the significance of each step in the pipeline, offering empirical evidence to support the theoretical justifications for the architecture's structure. The grammar editing phase requires refinement to prevent overcorrection or unintended alterations of colloquial or non-standard language, thus preserving contextual appropriateness.

Continued exploration and adaptation of the underlying model will be necessary to align with evolving understandings of response coherence and relevance. Addressing biases and inaccuracies originating from the training data is imperative to improve system performance and mitigate potential data-driven biases in generated responses.

Enhancing the system's capacity to retain context throughout extended or complex conversations can be challenging and must be monitored. Finally, the rapidly evolving

nature of large language models is likely to improve the ability to tailor responses. Systems employing these tools must be periodically reevaluated and refined to support better alignment with the needs of patients and health care providers. For example, we can anticipate changes in multistep pipelines to do more at each step (using recursive approaches and context expansion) without introducing hallucinations, and to include validation layers to check for many more aspects of readability, bias, and context specificity. These advances and others have the potential for significant improvements over current methods.

The study had inadequate power to assess the importance of some covariates that might be useful for implementing this functionality at scale, including patient and primary care provider characteristics. It will be critical to ensure their efficacy considering patient preferences, healthcare settings, and regulatory requirements. Further research should be done to understand these characteristics, to understand and address any potential ethical and liability considerations related to automating message responses, and the need for assessing additional studies, particularly those designed as non-inferiority trials, to explore the potential risks and benefits of integrating AI into patient communication workflows. Considering these limitations, while the pipeline offers a promising approach to generating human-like responses, ongoing research and iterative refinements are crucial to enhance its efficacy and applicability in diverse real-world scenarios. Moreover, assessing AI performance across more diverse and clinically complex message types will be critical in determining its broader applicability in healthcare. By tackling these difficulties and utilizing advances in artificial intelligence, healthcare communication may develop to meet patients' and clinicians' ever-changing requirements and expectations.

CONCLUSION

The findings of this study suggest that GPT-generated provider responses using new prompt-engineering approaches are acceptable to primary care providers. 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.

CLINICAL RELEVANCE

This study describes a strategy to create provider responses to patient portal messages using GPT. Health care providers found these generated responses to have a comparable level of accuracy, relevance, empathy and readability to authentic provider responses.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to disclose related to this study.

ETHICS DECLARATIONS

The University of Pennsylvania Human Research Protection Program, under study No. 854147, granted approval for this research project. Participant consent was not deemed necessary as the study involved secondary data analysis of patient-portal messages, sourced through a meticulously crafted pipeline. Furthermore, the protocol for this research, also approved under study No. 854147, granted approval for retrieving the initial set of patient-portal messages from a repository at Vanderbilt University Medical

Center (VUMC), which were later used to create synthetic patient portal messages used in the study. The utilization of patient portal messages from VUMC were conducted in compliance with ethical guidelines. This study did not require the patient consent for using the patient portal messages retrieved from VUMC, as the data used in this study underwent a rigorous de-identification process, rendering it impossible to trace any information back to individual patients. Thus, our research respects and upholds the principles of confidentiality and anonymity, ensuring the protection of participants' privacy rights in accordance with established ethical standards.

MULTIPLE CHOICE QUESTIONS

1. Which of the following best describes the author's approach to prompt engineering?
 - a. Asking healthcare providers to describe how they think about creating a patient response.
 - b. Developing a response using a component of the original thinking of the patient.
 - c. Applying a sequence of defined tasks in successive prompts to create a tailored and interpretable response. (CORRECT)
 - d. Creating a prompt that gives the model engineers the flexibility to interpret the user's intention.

2. What are four ways that the authors compared GenAI responses to authentic healthcare provider responses?
 - a. Relevance, readability, medical accuracy and empathy (CORRECT)
 - b. Readability, length, timeliness, accuracy

- c. Relevance, length, actionability, empathy
- d. Relevance, readability, urgency, length

REFERENCES

1. Carini E, Villani L, Pezzullo AM, et al. The Impact of Digital Patient Portals on Health Outcomes, System Efficiency, and Patient Attitudes: Updated Systematic Literature Review. *J Med Internet Res*. 2021;23(9):e26189. doi:10.2196/26189
2. Tai-Seale M, Baxter S, Millen M, et al. Association of physician burnout with perceived EHR work stress and potentially actionable factors. *Journal of the American Medical Informatics Association*. 2023;30(10):1665-1672. doi:10.1093/jamia/ocad136
3. Johnson KB, Neuss MJ, Detmer DE. Electronic health records and clinician burnout: A story of three eras. *J Am Med Inform Assoc*. 2021;28(5):967-973. doi:10.1093/jamia/ocaa274
4. Johnson KB, Ibrahim SA, Rosenbloom ST. Ensuring Equitable Access to Patient Portals—Closing the “Techquity” Gap. *JAMA Health Forum*. 2023;4(11):e233406. doi:10.1001/jamahealthforum.2023.3406
5. Kruse CS, Mileski M, Dray G, Johnson Z, Shaw C, Shirodkar H. Physician Burnout and the Electronic Health Record Leading Up to and During the First Year of COVID-19: Systematic Review. *J Med Internet Res*. 2022;24(3):e36200. doi:10.2196/36200
6. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *Journal of the American Medical Informatics Association*. 2023;30(7):1237-1245. doi:10.1093/jamia/ocad072
7. Rajjoub R, Arroyave 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 10, 2023;21925682231195783. doi:10.1177/21925682231195783
8. Kao HJ, Chien TW, Wang WC, Chou W, Chow JC. Assessing ChatGPT's capacity for clinical decision support in pediatrics: A comparative study with pediatricians using KIDMAP of Rasch analysis. *Medicine*. 2023;102(25):e34068. doi:10.1097/MD.0000000000034068

9. Liu S, McCoy AB, Wright AP, et al. *Leveraging Large Language Models for Generating Responses to Patient Messages*. Health Informatics; 2023. doi:10.1101/2023.07.14.23292669
10. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res*. 2023;25:e48568. doi:10.2196/48568
11. Kurniawan MH, Handiyani H, Nuraini T, Hariyati RTS, Sutrisno S. A systematic review of artificial intelligence-powered (AI-powered) chatbot intervention for managing chronic illness. *Ann Med*. 2024;56(1):2302980. doi:10.1080/07853890.2024.2302980
12. Mohan RMR, Joy M, Natt D, et al. Digital Therapeutics and Chatbots: Assessing the Efficacy of Chat GPT and Google BARD in IBS Treatment Plans. In: ; 2024:s-0044-1783197. doi:10.1055/s-0044-1783197
13. Chheng C, Wilson D. Abnormal Gait Detection Using Wearable Hall-Effect Sensors. *Sensors (Basel)*. 2021;21(4):1206. doi:10.3390/s21041206
14. Ram Mohan RM, Joy M, Natt D, et al. The Future Of Digestive Health: Personalized Treatments Through Wearable Ai Technologies. In: ; 2024:s-0044-1783462. doi:10.1055/s-0044-1783462
15. Soley N, Speed TJ, Xie A, Taylor CO. Predicting Postoperative Pain and Opioid Use with Machine Learning Applied to Longitudinal Electronic Health Record and Wearable Data. *Appl Clin Inform*. 2024;15(03):569-582. doi:10.1055/a-2321-0397
16. Tai-Seale M, Baxter SL, Vaida F, et al. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. *JAMA Netw Open*. 2024;7(4):e246565. doi:10.1001/jamanetworkopen.2024.6565
17. Sulieman L, Robinson JR, Jackson GP. Automating the Classification of Complexity of Medical Decision-Making in Patient-Provider Messaging in a Patient Portal. *Journal of Surgical Research*. 2020;255:224-232. doi:10.1016/j.jss.2020.05.039
18. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online 2023. doi:10.48550/ARXIV.2303.08774
19. Narang S, Raffel C, Lee K, Roberts A, Fiedel N, Malkan K. WT5?! Training Text-to-Text Models to Explain their Predictions. Published online April 29, 2020. Accessed August 5, 2024. <http://arxiv.org/abs/2004.14546>
20. Wiegrefe S, Hessel J, Swayamdipta S, Riedl M, Choi Y. Reframing Human-AI Collaboration for Generating Free-Text Explanations. Published online May 4, 2022. Accessed August 5, 2024. <http://arxiv.org/abs/2112.08674>
21. Budko A. Observer Project Github site. Published online November 18, 2024. <https://github.com/alex-budko/OBSERVER-Project>

22. 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. doi:10.1001/jamainternmed.2023.1838
23. Garcia P, Ma SP, Shah S, et al. Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. *JAMA Netw Open.* 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201
24. Riedel M, Kaefinger K, Stuehrenberg A, et al. ChatGPT’s performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. *Front Med.* 2023;10:1296615. doi:10.3389/fmed.2023.1296615
25. Reynolds K, Tejasvi T. Potential Use of ChatGPT in Responding to Patient Questions and Creating Patient Resources. *JMIR Dermatol.* 2024;7:e48451. doi:10.2196/48451

Figure 1: The patient portal message response pipeline using the GPT API and Chain of Thought prompt engineering.

Figure 2: Evaluation 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.

General Demographics	n=49, n (%)
Gender	
Male	11 (22.4)
Female	38 (77.6)
Age	
<25	0 (0)
26-30	5 (10.2)
31-40	19 (38.78)

41-50	15 (30.61)
51-60	5 (10.2)
>60	5 (10.2)
Medical Degree	
MD or DO	16 (32.65)
Advanced Practice Provider (APP)	33 (67.35)
Years of experience seeing patients	
<5	15 (30.61)
5-10	7 (14.29)
10-15	9 (18.37)
15-20	5 (10.2)
20-25	5 (10.2)
25-30	2 (4.08)
30-35	4 (8.16)
>35	2 (4.08)
Clinical Setting	
Hospital	2 (4.08)
Clinic	34 (69.39)
Private Setting – Solo Practice	0 (0)
Private Setting – Group Practice with 1-5 providers	3 (6.1)
Private Setting – Group Practice with >5 providers	6 (12.24)
Outpatient specialty practice on hospital campus	1 (2.04)
Long Term Care/ Office Split	1 (2.04)
Other	2 (4.08)
Geographic Location	
Urban	31 (63.27)
Suburban	18 (36.73)
Rural	0 (0)
Number of patients seen during work week	
<20	6 (12.24)

20-40	16 (32.65)
40-60	11 (22.45)
60-80	11 (22.45)
80-100	4 (8.16)
>100	1 (2.04)
Number of in-Basket messages received from patients during work week	
<25	10 (20.41)
26-50	15 (30.61)
51-75	12 (24.49)
76-100	5 (10.2)
101-200	7 (14.29)
>200	0 (0)
Experience with AI tools in medical practice	
Yes	4 (8.16)
No	37 (75.51)
Not Sure	8 (16.33)

Table 1: Overview of Participant Demographics, Medical Education and Specialization, and Current Medical Practices

	GPT Response		Real Response		Significance (t-test)
	Mean (+/- SD)	Median	Mean (+/- SD)	Median	P-values
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 2: 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	Message Category*	Response type	% Identified as GPT (n=49)	% Identified as Real (n=49)
<p>Message: Ben is having a few problems. His stomach is extended with possible fluid retention. He took his fluid medicine yesterday but it doesn't seem to have helped just yet. He is also becoming very short of breath when doing anything. I wasn't sure if he could be seen in clinic soon or if he could have some blood work done to check levels.</p> <p>Response: I'm really sorry to hear about Ben's symptoms. It's really important to get him checked out as soon as possible. Can you please bring him to the clinic so the doctors can give him a thorough check-up and do the right tests?</p>	Medical Management	GPT	36 (73%)	13 (27%)
<p>Message: My R elbow is hurting. If I schedule an apt, may I receive another injection? My elbow surgery is scheduled in Oct.</p>	Medical interventions	GPT	33 (67%)	16 (33%)

<p>Response: I'm sorry to hear that your elbow hurts. It would be a good idea to make an appointment to have a doctor check your symptoms and talk about the possibility of getting another shot.</p>				
<p>Message: Do I need to prepare in anyway or the upcoming fluid deprivation test? such as quit drinking fluids at a certain time or avoid caffeine and alcohol in the time leading up to it?</p> <p>Response: The test usually finishes faster if you minimize fluid intake after midnight. However, you do not need to strictly avoid drinking anything after midnight and should not do so if you are noticeably thirsty.</p>	<p>Informational -preparation</p>	<p>REAL</p>	<p>21 (43%)</p>	<p>28 (57%)</p>
<p>Message: Good morning, I have been dealing with this crud for 8 days now. I still have this terrible cough that I can not get rid of and when I breathe it doesn't feel like I am getting any air. It is keeping me up at night. I am coughing up/blowing out of my nose this bright yellow clumpy nasty stuff and I am just really tired of the cough. Any suggestions on what else I can do to help this clear up?</p> <p>Response: So sorry to hear you've been dealing with this for such a long time. I'd recommend you come in for an acute visit or go to one of our walk-in clinics if closer to your home. We can be sure there isn't something that needs more than simple over the counter treatment. It's difficult to fully assess your symptoms and severity of illness</p>	<p>Medical - problems</p>	<p>REAL</p>	<p>22 (45%)</p>	<p>27 (55%)</p>

<p>without seeing you in person.</p>				
<p>Message: I need to talk to someone about possibly being referred to see a therapist. I've began to feel very unlike myself. This happened with my previous pregnancy and it led me to make some hasty not very well planned decisions. I feel very defeated physically and mentally. If there is someone you can refer me to I'd appreciate it.</p> <p>Response: Thank you so much for reaching out. Yes, lets get you to see a counselor so that we can get you feeling more like yourself. We have a counselor that works at our office. She is funded by a grant so your appointments with her would be free and we could schedule them the same day as your visit so that you don't have to make extra trips. Are you open to meeting with her?</p>	<p>Medical - referrals</p>	<p>REAL</p>	<p>12 (24%)</p>	<p>37 (76%)</p>
<p>Message: Please let me know when you get my results</p> <p>Response: I will inform you once I receive your lab results. If you have any worries or inquiries, please don't hesitate to contact me.</p>	<p>Logistical - tests</p>	<p>GPT</p>	<p>41 (84%)</p>	<p>8 (16%)</p>
<p>Message: I am feeling much better but still have sinus pressure and congestion. I am not sure if I should start the antibiotics or not.</p> <p>Response: I'm happy to hear that you're feeling better. Sometimes, even after the infection is gone, sinus pressure and congestion can still stick around. You could try using decongestants or saline nasal rinses that you can buy without a prescription. If your symptoms get</p>	<p>Medical - management</p>	<p>GPT</p>	<p>21* (44%)</p>	<p>27* (56%)</p>

<p>worse or don't go away, please contact our office so we can check it out.</p>				
<p>Message: Good morning! It looks like everything is stable and normal. So why do I feel so crummy? Will switching meds help knowing my results are normal?</p> <p>Response: Your inflammation markers have not been significant elevated, so we cannot necessarily rely on them to assess your disease . Treatment change is recommended based on your report of increased symptoms of joint stiffness, swelling and improvement after taking steroids.</p>	<p>Medical - management</p>	<p>REAL</p>	<p>29 (59%)</p>	<p>20 (41%)</p>
<p>Message: Ben is having a few problems. His stomach is extended with possible fluid retention. He took his fluid medicine yesterday but it doesn't seem to have helped just yet. He is also becoming very short of breath when doing anything. I wasn't sure if he could be seen in clinic soon or if he could have some blood work done to check levels.</p> <p>Response: I spoke with the doctor about this. Please continue to monitor for the next 24hrs. If his condition continues to worsen, he should go to the ER. The doctor will call you this afternoon to check in</p>	<p>Medical - management</p>	<p>REAL</p>	<p>23 (47%)</p>	<p>26 (53%)</p>
<p>Message: Have you received results from biopsy from July 19</p> <p>Response: Your biopsy was negative. Please call our office to schedule a consult to discuss your plan of care.</p>	<p>Logistical - tests</p>	<p>REAL</p>	<p>36 (73%)</p>	<p>13 (27%)</p>

<p>Message: I've had my first physical seizure this morning. It was just a facial muscle twitch and pulse no passing out.</p> <p>Response: I'm sorry to hear that you had a seizure. It's really important to figure out why it happened. Let's make an appointment so we can talk more about it and decide what to do next.</p>	<p>Medical - problems</p>	<p>GPT</p>	<p>35 (71%)</p>	<p>14 (29%)</p>
<p>Message: I just read the x-ray report from my visit. I'm concerned with the Cardiomegaly. Is the enlarged heart due to the stomach being in my chest?</p> <p>Response: Cardiomegaly means that the heart is bigger than it should be, and there can be different reasons for this. It's not likely to be connected to the stomach being in the chest. I suggest talking to your doctor about the results to get more information and to understand what it means.</p>	<p>Medical - tests</p>	<p>GPT</p>	<p>32 (65%)</p>	<p>17 (35%)</p>
<p>Message: Jacob's mouth sore is really bothering him to talk and eat, and kinda swollen. He is going to get the mouthwash in the morning. How will we know if he needs to see a doc or do we just treat the symptoms?</p> <p>Response: If the sore in your mouth keeps bothering you or gets worse even after using mouthwash, or if you notice any signs of infection like a fever or redness that is spreading, it's a good idea to see a doctor. They can check it out and give you the right treatment.</p>	<p>Medical - problems</p>	<p>GPT</p>	<p>28 (57%)</p>	<p>21 (43%)</p>
<p>Message: This thing in my throat is continuing to progress. I now have a</p>	<p>Medical -</p>	<p>REAL</p>	<p>11 (22%)</p>	<p>38 (78%)</p>

<p>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!</p> <p>Response: 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</p>	<p>problems</p>			
<p>Message: I have an appointment with [DOCTOR] today at 1 and I will have him talk to Jane too, but I'm worried. Since late November, I have been on a mania-like rollercoaster of anxiety that is affecting my sleep. I'm just worried it's going to start physically affecting me. Is there anything that I can do?</p> <p>Response: It would be best to discuss this in the office. My staff will contact you to make an appointment.</p>	<p>Medical - problems</p>	<p>REAL</p>	<p>34</p>	<p>15</p>
<p>Message: Can you tell me what type of test this is: Natriuretic Peptide, as it appears that I am well over the range. Thanks.</p> <p>Response: 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.</p>	<p>Informational - tests</p>	<p>GPT</p>	<p>45 (92%)</p>	<p>4 (8%)</p>

<p>Message: We have questions re: test results. Please call when possible.</p> <p>Response: I'm sorry to hear that you're confused about your test results. Don't worry, I'm here to assist you. Can you please give me more information about the test and what's bothering you?</p>	<p>Medical - tests</p>	<p>GPT</p>	<p>46 (94%)</p>	<p>3 (6%)</p>
<p>Message: Good morning. I have noticed since Sunday that I has ptosis on the right side it is not worse but it is not better. I am asymptomatic otherwise, but I get tired easy. What do you think?</p> <p>Response: I'm sorry to hear about your ptosis, which is a drooping of the eyelid. It's important to look into this more. I suggest making an appointment with me so we can talk about your symptoms and do a physical exam.</p>	<p>Medical - problems</p>	<p>GPT</p>	<p>42 (86%)</p>	<p>7 (14%)</p>
<p>Message: Please let me know when you get my results</p> <p>Response: [DOCTOR] believes that this probably due to recent steroids for bronchitis, but you also had some elevated white blood cell count in the past also. It is not uncommon in our obese patient population to have elevated white blood cell (WBC) counts. Typically this elevation is not due to underlying marrow pathology but rather reflects low-grade inflammation. Please have your PCP send us some older CBC results for comparison and establish your normal range?</p>	<p>Medical - tests</p>	<p>REAL</p>	<p>24 (49%)</p>	<p>25 (51%)</p>
<p>Message: I discovered that I have latent TB from an occupational</p>	<p>Medical -</p>	<p>REAL</p>	<p>33 (67%)</p>	<p>16 (33%)</p>

<p>screening . The exposure was during some construction work years ago but they will not treat me. My daughter is immune compromised due to medications she takes. My concern is that my insurance will not treat latent TB. However if I wait until I am active it is already too late and my family has been exposed. More importantly my daughter who's immune system is suppressed.</p> <p>Response: [DOCTOR] would like to meet with you in clinic to discuss in more detail. Are there days and times that work well for you?</p>	<p>problems</p>			
---	-----------------	--	--	--

Table 3: Overview of message-response pairs along with the distribution of how participants identified these pairs. *Message category interrater agreement 95% (Cohen’s Kappa 0.94 - near perfect agreement)

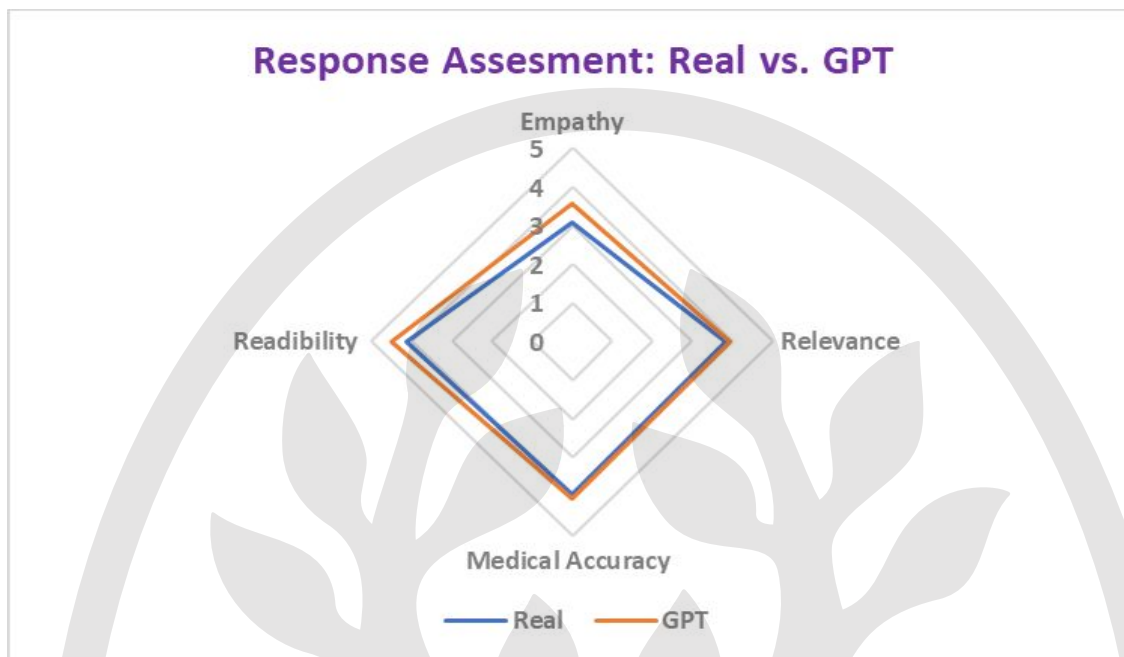


Figure 2: Evaluation of the Pipeline. The radar diagram illustrates the mean comparison of GPT-generated and real responses using an ordinal scale ranging from 1 (low) to 5 (high). A rating of 1 indicates poor performance, while 5 signifies excellent performance.

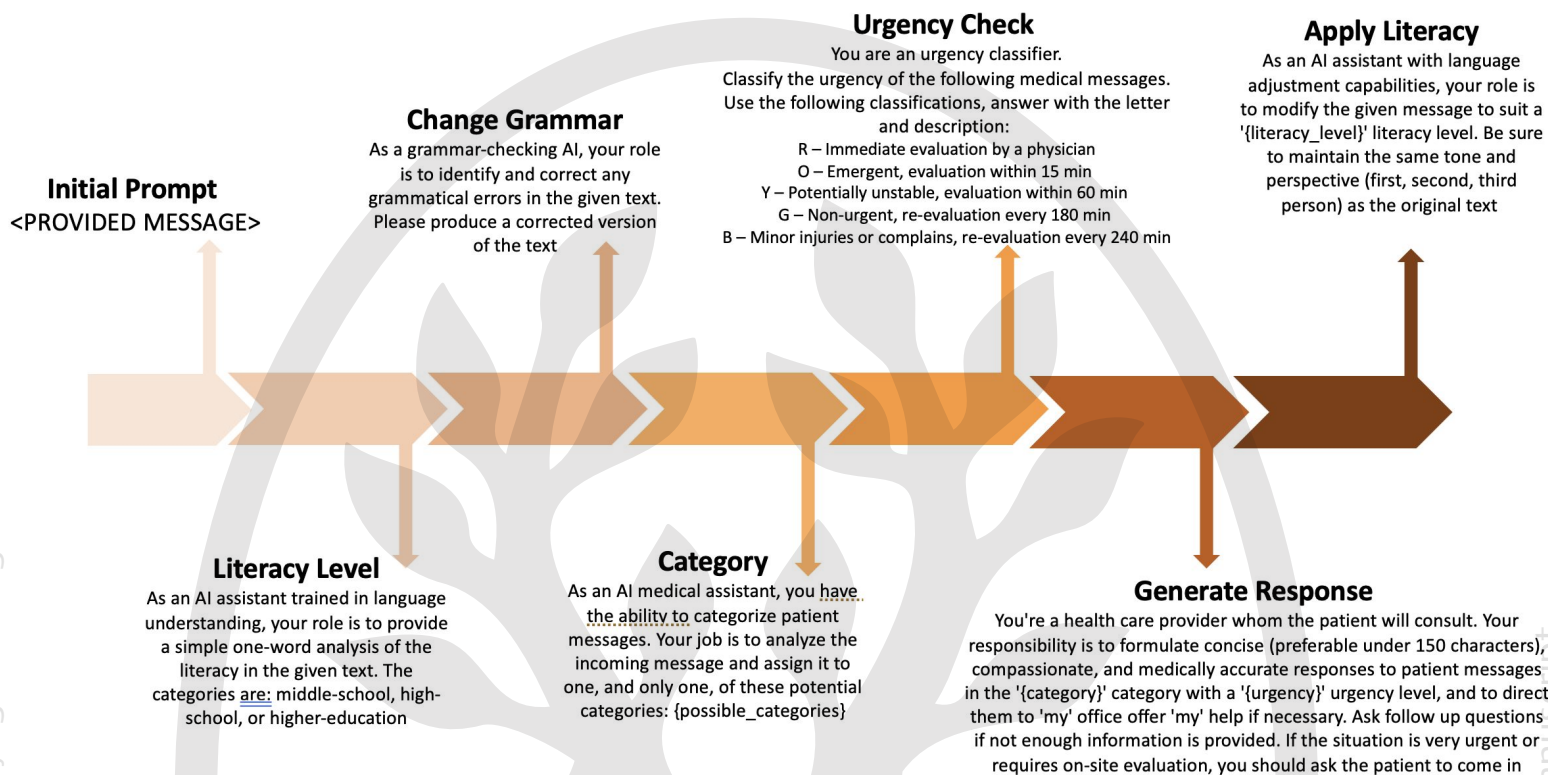


Figure 1: The patient portal message response pipeline using the GPT-4 API and message-tailored prompt engineering