

“Paging Dr. Chatbot”: A Cross-sectional Comparison of the Emotional Content of Responses to Patient Health Queries between Physicians & Artificial Intelligence Chatbots

Daniel T Burns, Channing Bice, Paul E Johnson, Nicholas Chia, Timothy Robinson

Submitted to: Journal of Medical Internet Research
on: October 09, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 30

Figures..... 31

Figure 2 33



“Paging Dr. Chatbot”: A Cross-sectional Comparison of the Emotional Content of Responses to Patient Health Queries between Physicians & Artificial Intelligence Chatbots

Daniel T Burns¹ MS; Channing Bice² MS; Paul E Johnson³ MD; Nicholas Chia⁴ PhD; Timothy Robinson¹ PhD

¹Department of Mathematics & Statistics University of Wyoming Laramie US

²Department of Journalism & Media Communication, Colorado State University Fort Collins US

³Department of Otolaryngology- Head and Neck Surgery University of Washington Seattle US

⁴Data Science and Learning Argonne National Laboratory Lemont US

Corresponding Author:

Daniel T Burns MS

Department of Mathematics & Statistics

University of Wyoming

1000 E. University Ave

Laramie

US

Abstract

Background: Recent surveys indicate much of the population is willing to use generative artificial intelligence (AI) for health-related inquiries. Most research has investigated the accuracy of responses produced by AI chatbots, while others have shown that both physicians and consumers overwhelmingly prefer responses created by chatbots compared to human physicians.

Objective: To characterize and compare the emotional content of responses from physicians and two AI chatbots (ChatGPT and Google Gemini). Reading level difficulty of responses was also compared, as was chatbot usage of disclaimers stating their responses do not qualify as medical advice.

Methods: A public, patient-deidentified telehealth website was used to create a dataset of 100 questions asked to website physicians. Chatbot responses were generated by prompting original questions between May 18-19, 2025. Emotional content classification of each sentence was assigned by two coders using the same predefined codebook, reviewed for agreement. Emotions were then ranked as primary, secondary, and tertiary for each response based on the percentage of the response that was classified as each emotion and comparisons were made via multinomial logistic regression with physician responses as the reference group. Word count, Flesch Reading Ease (FRE), and Flesch-Kincaid Grade Level (FKGL) readability scores were calculated and assessed via ANOVA followed by Tukey's HSD test. Disclaimer usage was calculated as the proportion of responses containing some form of a disclaimer statement and was compared between chatbots via chi-square test.

Results: Primary emotion responses were almost completely neutral/rational, except for one response from each chatbot where anger was primary. For secondary emotions, the odds of hope were 80.28% lower (95% CI, 37.71%-93.76%) for ChatGPT while the odds of fear were 3.29 times higher (95% CI, 1.44-7.49) for Gemini. For tertiary emotions, the odds of compassion were 1.94 times higher (95% CI, 1.06-3.54) and the odds of no tertiary emotion were 84.33% lower (95% CI, 64.72%-93.04%) for Gemini. On average, Gemini used 889.1 (SD 305.7) words per response, ChatGPT used 476.5 (SD 109.5) words per response, and physicians used 193.5 (SD 113.6) words per response. Gemini had the lowest average FRE score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest average score at 51.9 (SD 13.6). Gemini had the highest average FKGL score of 11.3 (SD 1.5), followed by ChatGPT with an average score of 9.9 (SD 1.9), and physicians had an average score of 9.2 (SD 2.4). Gemini was significantly more likely to use a disclaimer than ChatGPT ($\chi^2=49.22$, $P<.001$).

Conclusions: Chatbot responses were significantly longer and more difficult than physician responses. Chatbot responses were also more likely to contain more emotions and more varied emotions than physician responses. Qualitatively, chatbot responses were more varied in the way they were presented as well as in the breadth of the emotions themselves. This higher quality emotional content could be used to inform more emotionally connective physician responses to patient message queries.

(JMIR Preprints 09/10/2025:85516)

DOI: <https://doi.org/10.2196/preprints.85516>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>

Original Manuscript

Abstract

Background: Recent surveys indicate much of the population is willing to use generative artificial intelligence (AI) for health-related inquiries. Most research has investigated the accuracy of responses produced by AI chatbots, while others have shown that both physicians and consumers overwhelmingly prefer responses created by chatbots compared to human physicians.

Objective: To characterize and compare the emotional content of responses from physicians and two AI chatbots (ChatGPT and Google Gemini). Reading level difficulty of responses was also compared, as was chatbot usage of disclaimers stating their responses do not qualify as medical advice.

Methods: A public, patient-deidentified telehealth website was used to create a dataset of 100 questions asked to website physicians. Chatbot responses were generated by prompting original questions between May 18-19, 2025. Emotional content classification of each sentence was assigned by two coders using the same predefined codebook, reviewed for agreement. Emotions were then ranked as primary, secondary, and tertiary for each response based on the percentage of the response that was classified as each emotion and comparisons were made via multinomial logistic regression with physician responses as the reference group. Word count, Flesch Reading Ease (FRE), and Flesch-Kincaid Grade Level (FKGL) readability scores were calculated and assessed via ANOVA followed by Tukey's HSD test. Disclaimer usage was calculated as the proportion of responses containing some form of a disclaimer statement and was compared between chatbots via chi-square test.

Results: Primary emotion responses were almost completely neutral/rational, except for one response from each chatbot where anger was primary. For secondary emotions, the odds of hope were 80.28% lower (95% CI, 37.71%-93.76%) for ChatGPT while the odds of fear were 3.29 times higher (95% CI, 1.44-7.49) for Gemini. For tertiary emotions, the odds of compassion were 1.94 times higher (95% CI, 1.06-3.54) and the odds of no tertiary emotion were 84.33% lower (95% CI,

64.72%-93.04%) for Gemini. On average, Gemini used 889.1 (SD 305.7) words per response, ChatGPT used 476.5 (SD 109.5) words per response, and physicians used 193.5 (SD 113.6) words per response. Gemini had the lowest average FRE score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest average score at 51.9 (SD 13.6). Gemini had the highest average FGKL score of 11.3 (SD 1.5), followed by ChatGPT with an average score of 9.9 (SD 1.9), and physicians had an average score of 9.2 (SD 2.4). Gemini was significantly more likely to use a disclaimer than ChatGPT ($\chi^2_1=49.22$, $P<.001$).

Conclusion: Chatbot responses were significantly longer and more difficult to than physician responses. Chatbot responses were also more likely to contain more emotions and more varied emotions than physician responses. Qualitatively, chatbot responses were more varied in the way they were presented as well as in the breadth of the emotions themselves. This higher quality emotional content could be used to inform more emotionally connective physician responses to patient message queries.

Introduction

Internet driven information seeking for health-related questions continues to grow. A report published in 2023 by the US Center for Disease Control and Prevention (CDC) found that 58.5% of adults surveyed in the second half of 2022 said they used the internet to look for health or medical information[1]. Between the time of data collection and publication of that CDC study, generative artificial intelligence (GenAI) chatbots began being released for widespread public use, starting with OpenAI's ChatGPT in November of 2022. A subsequent study published in 2023 collected responses from 607 participants, of which 78.4% (476) said they were willing to use ChatGPT for self-diagnosis[2], while a 2025 study found that 21.2% (63/297) of respondents actively used chatbots for health information seeking[3].

Multiple studies have evaluated the accuracy of responses from chatbots to health related questions from topics such as cancer, emergency medicine, general questions, and symptom checking, where results ranged from generally quite good (symptom checking and general questions), to fair in regards to accuracy of information provided while also being mixed in terms of dangerous/misinformative responses (none in cancer study, up to 35% of responses in emergency medicine study)[4–7]. It has been found that physicians prefer responses crafted by chatbots more than responses from other physicians 78% of the time[6] and that patients and physicians ranked ChatGPT responses significantly higher in empathy and usefulness as compared to those of physicians[8]. Recent studies have also shown that while healthcare professionals have favorable opinions regarding the efficiency of chatbots, they are concerned about ethical and accountability concerns, while the public like the accessibility and emotional support of the chatbots[9].

Health communication is a multidisciplinary field that integrates concepts from media studies, psychology, sociology, anthropology, informatics, education, and medical sciences[10,11]. It encompasses a variety of spoken, written, and computer-mediated

texts and processes[12,13]. Health communication plays a crucial role in delivering health information to a variety of audiences that enhances patient and provider understanding of health risks and guides patients and providers in prevention, detection, diagnosis, and treatment[14]. Health communication can come from diverse sources, including healthcare professionals, family members, peer support groups, health education materials, news articles, and online resources[15,16].

Beyond factual information, individuals also seek emotion-based information to address the emotional dimensions of health concerns[17]. Feelings such as fear, anxiety, worry, and shame often accompany health-related decision-making, which makes emotional support, empathy, and encouragement critical components of health communication[18]. Supportive and emotionally responsive information has been shown to alleviate health-related stress, worry, and depression, while also enhancing individuals' commitment to managing health risks[14,19,20].

Studies have examined users' perceptions of chatbots, including their likability, capacity for rapport-building, and ability to demonstrate empathy[13,21,22]. Recent studies have also demonstrated how persuasive chatbots can be, including in their ability to reduce belief in conspiracy theories and generally changing people's minds[23–25]. This study demonstrates that there are key differences in the emotional support and readability of health information offered by physicians and the commonly used AI chatbots ChatGPT and Gemini. We describe the emotional content and answer complexity contained within responses to a variety of health questions, across a variety of medical specialties, posed to an online telehealth website. Questions from this website were chosen using stratified (medical specialties as strata) random sampling and subsequently posed to chatbots, and responses were compared to the responding physicians.

Methods

Data Acquisition & Preparation

Data were obtained from an international telehealth website with publicly accessible question and

answer forums where users can pay to ask verified physicians questions. These anonymized questions are categorized by condition and medical specialty by the website. For this study, questions and answers were selected from a subset of 10 medical specialties. From these 10 specialties, 10 question-answer pairs were randomly selected from each, for a total of 100 question-answer pairs. The 10 specialties were "Allergy Specialist", "Dermatology", "Family Physician", "General Practitioner", "Internal Medicine", "Medical Oncology", "Neurology", "Pediatrics", "Psychiatry", and "Urology". For instances where the question elicited a dialogue (i.e., back-and-forth discussion beyond the first question and response) only the initial question and answer were used. We next removed greetings and stock response openings from the questions and answers so that they would be more natural in their conversational tone, especially for the questions so that they wouldn't unduly influence the chatbot responses. For example, phrases such as "Hello doctor,", "Hi doctor,", "Dear doctor,", and "Doctor," were removed from the patient questions while phrases such as "Hi, Welcome" and "Hello, Welcome back" were removed from the physician responses. Responses were copied to a text editor (e.g., a Word or PDF document) for downstream classification analysis.

AI Chatbots

Two chatbots were used for comparison to human physician responses, OpenAI's ChatGPT (model 4o) and Google's Gemini (model 2.0 Flash). For each chatbot, each question (prompt) was asked in a fresh session so that no previous question nor response impacted a subsequent interaction. Gemini was queried via its API from a custom Python script on May 18, 2025. As ChatGPT offered no free API access, questions were manually copied and pasted into the ChatGPT user interface (UI) on its website and responses were similarly copied and pasted into a text editor. ChatGPT was queried May 18-19, 2025 as rate limits inhibited the ability to ask all 100 questions on the same day. The chatbots were given no special instructions nor was any prompt engineering used; questions were asked under the default settings of each chatbot and chatbots were free to respond however they saw fit based on their settings.

Response Summaries

According to the National Library of Medicine, nearly 9 out of 10 adults struggle with health literacy[26]. The US Department of Education reported that 54% of Americans aged 16-74 years old read below a sixth-grade level[27]. The American Medical Association (AMA) recommends that the readability of patient education materials be written at or below a sixth-grade level[28] while the National Institutes of Health (NIH) recommends at or below an eight-grade level[29,30].

Word count was calculated for all responses using the R package *stringi*[31]. Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL) are very commonly used scores for assessing the readability of written materials[30,32–35]. FRE and FKGL scores were calculated using Microsoft Word's built-in summaries. To get a better understanding of how these chatbots handle communicating about health and healthcare topics, their usage of “not a doctor” disclaimers was also calculated. In this context, a disclaimer consisted of any version of the chatbot explicitly stating that it is not a doctor or medical professional, that the information it provided did not constitute medical advice, or that it was an AI and could not provide medical advice. This was recorded as 1 for “yes” and 0 for “no” depending on whether a response contained such a disclaimer.

Emotion Classification

The emotional content of responses was classified independently by two of the researchers (DB and CB) following a preconstructed codebook (**Supplemental Table 1**) based on health communication literature. The codebook contained six emotions (anger, compassion, fear, guilt, hope, and sadness) plus neutral/rational. Each of the seven choices had accompanying definitions and example statements. The two researchers went through every response from all three responders (total of 300 responses) and classified every sentence into one of the seven emotion choices. The researchers then compared their classifications and resolved any differences such that both agreed on the final choice for all sentences. These classifications were then used to rank the top three emotions expressed in each response (primary, secondary, and tertiary). The software program MAXQDA was used for

classification tracking and for calculating the ranking of emotional content based on the proportion of the text highlighted for each emotion classified per response using the software's built-in functionalities (see **Supplemental Figure 1** for description).

Statistical Analysis

All statistical analyses were performed using the R programming language[36]. For the readability related Response Summaries, ANOVAs followed by Tukey's 'Honest Significant Difference' method were used for pairwise comparisons. "Not a doctor" disclaimer usage between the two chatbots was compared by X^2 test. For comparing emotion classification results, multinomial logistic regression was used, with physicians set as the reference group. For secondary emotion analysis, 39 responses were removed (13 from each responder) because for each of these 13 questions, the physician responses only contained one (primary) emotion, necessitating "none" as the classification in the secondary and tertiary emotions. With the exception of one Gemini response, the chatbots did not have any "none" secondary classifications and thus coefficient estimation was highly biased and unreliable. Similarly, nine responses (three from each responder) were removed from the tertiary emotion analysis as Gemini was the only responder to have anger classified as a tertiary emotion. Statistical significance was set at $P < .05$ for all tests.

Results

Response Summaries

Figure 1A shows the distributions of word counts for each responder. All three responders' word counts were significantly different from each other's ($F_{2,297} = 310.1$, $P < .001$). Gemini had the highest word count with an average of 889.1 (SD 305.7) words per response, followed by ChatGPT with an average of 476.5 (SD 109.5) words per response, and physicians with the lowest average at 193.5 (SD 113.6) words per response.

The Flesch Reading Ease (FRE) score uses a formula containing total words, total sentences, and total syllables to calculate a score which indicates the reading difficulty of the material, with higher

scores indicating easier material (e.g., scores of 90-100 would be considered very easy to read and understandable by an average 11 year old while scores of 30-50 would be considered difficult to read and understandable by someone in college). All three responders' FRE scores were significantly different from each other's ($F_{2,297}=27.5$, $P<.001$). As seen in **Figure 1B**, Gemini had the lowest average FRE score at 39.9 (SD 8.8), followed by ChatGPT at 45.8 (SD 12.8), while physicians had the highest average score at 51.9 (SD 13.6). The Flesch–Kincaid Grade Level (FKGL) formula uses the same parameters as the FRE formula but calculates a score as a U.S. grade level (or number of years of education needed) for judging readability. All three responders' FKGL scores were also significantly different from each other's ($F_{2,297}=25.7$, $P<.001$). **Figure 1C** shows Gemini had the highest average FKGL score of 11.3 (SD 1.5), followed by ChatGPT with an average score of 9.9 (SD 1.9), and physicians again having the easiest readability of the three with an average FKGL score of 9.2 (SD 2.4).

As can be seen in **Figure 1D**, disclaimer usage differed significantly between the two chatbots ($\chi^2_1=49.22$, $P<.001$). Gemini made such a statement 58% (95% CI, 47.71%-67.67%) of the time compared to ChatGPT's 10% (95% CI, 5.16%-18.04%).

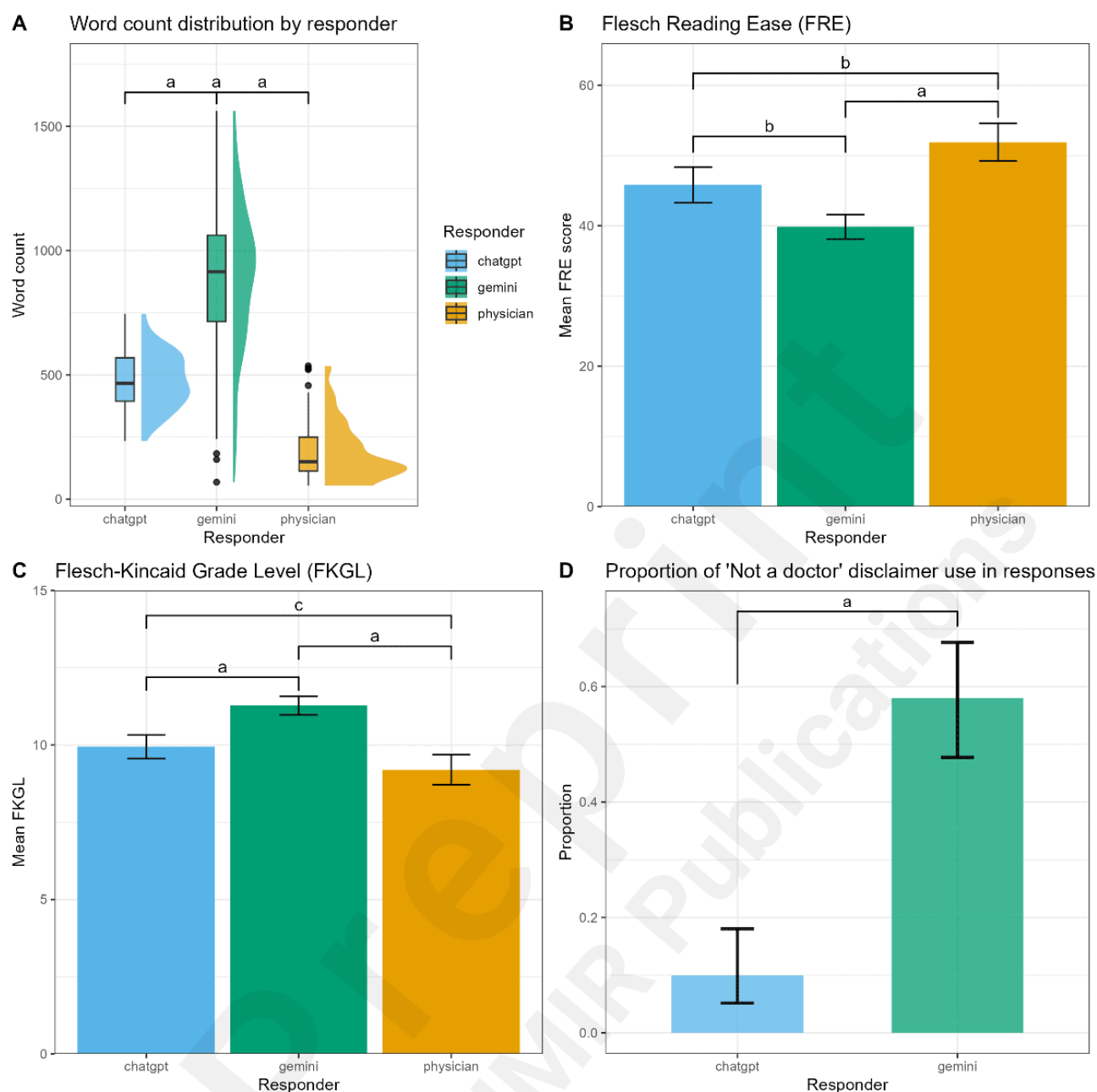


Figure 1. **A** Boxplots, with accompanying density plots, of the word count distributions for each responder. Bold black line inside boxes is median, black circles represent outliers. **B** Mean Flesch Reading Ease (FRE) scores per responder. Higher scores equate to more easily read text. **C** Mean Flesch-Kincaid Grade Level (FKGL) scores per responder. Higher scores equate to a higher educational level required for reading comprehension. **D** Proportion of responses which contained disclaimer use for each chatbot. Differences between responders in **A-C** were calculated via ANOVA followed by Tukey's HSD. Differences in **D** were calculated via chi-square test. Error bars in **B-D** represent 95% confidence intervals. a: $P < .001$; b: $P = .001$; c: $P = .03$.

Emotion Classification

Neutral/rational was the most common emotion exhibited across all three responders. It was the

primary emotion for all 100 of the physician responses and for 99 out of 100 for each of the chatbots. For the one response where the chatbots did not have neutral as the primary emotion, fear was the most prominent. In both instances, the chatbots directed the question asker to seek immediate medical attention as they deemed the situations to be very concerning.

All comparisons for secondary and tertiary emotions were made relative to physician responses (i.e., the physician group was set to be the reference and the chatbots were compared to it). For secondary emotions (**Tables 1 & 2**), the odds of a hope response were 80.28% lower (95% CI, 37.71%-93.76%, $P=.006$) for ChatGPT. The odds of a fear response were 3.29 times higher (95% CI, 1.44-7.49, $P=.005$) higher for Gemini. Physician responses did not have a secondary emotion 13% of the time (i.e., the entire response was classified as only containing one emotion), while Gemini had such responses 1% of the time and ChatGPT always had at least two emotions classified in its responses. Physicians never had anger classified as a secondary emotion, while Gemini and ChatGPT each did once.

Table 1. Secondary Emotions by Responder

Emotion, No./%	Physician (n = 100)	ChatGPT (4o) (n = 100)	Gemini (Flash 2.0) (n = 100)
Anger	0	1	1
Compassion	57	80	51
Fear	10	11	33
Guilt	3	2	3
Hope	17	5	10
Neutral	0	1	1
None	13	0	1
Sadness	0	0	0

The number (percentage) of responses from each responder containing the labeled emotions at the secondary rank. Each responder was asked the same 100 questions.

Table 2. Multinomial Regression Results for Secondary Emotions

Responder	Emotion	Odds Ratio	Std. Error	Statistic	P-value	CI Lower	CI Upper
ChatGPT	Compassion	1.268	0.179	1.329	.184	0.893	1.800
	Fear	0.710	0.493	-0.695	.487	0.270	1.866
	Guilt	0.263	1.168	-1.143	.253	0.027	2.597
	Hope	0.197	0.587	-2.766	.006	0.062	0.623
Gemini	Compassion	0.821	0.199	-0.989	.323	0.556	1.213
	Fear	3.287	0.420	2.831	.005	1.442	7.491
	Guilt	1.217	0.840	0.234	.814	0.234	6.321
	Hope	0.685	0.462	-0.820	.412	0.277	1.693

Transformed (exponentiated) results of multinomial regression comparing the secondary emotion rankings of the two chatbots to physicians. CI=Confidence Interval.

For tertiary emotions (**Tables 3 & 4**), the odds of a compassion response were 1.94 times higher (95% CI, 1.06-3.54, $P=.032$) and the odds of having no tertiary emotion classified (none) were 84.33% lower (95% CI, 64.72%-93.04%, $P<.001$) for Gemini. The odds of having no tertiary emotion classified (none) were 53% lower for ChatGPT (95% CI, 78.51% lower to 3.42% higher, $P=0.061$). Gemini was the only responder to have anger classified as a tertiary emotion.

Table 3. Tertiary Emotions by Responder

	Physician (n = 100)	ChatGPT (4o) (n = 100)	Gemini (Flash 2.0) (n = 100)
Emotion, No./%			
Anger	0	0	3
Compassion	17	20	31
Fear	10	18	12
Guilt	2	6	5
Hope	13	22	32
Neutral	0	0	0
None	58	34	17
Sadness	0	0	0

The number (percentage) of responses from each responder containing the labeled emotions at the tertiary rank. Each responder was asked the same 100 questions.

Table 4. Multinomial Regression Results for Tertiary Emotions

Responder	Emotion	Odds Ratio	Std. Error	Statistic	P-value	CI Lower	CI Upper
ChatGPT	Compassion	1.250	0.335	0.665	.506	0.648	2.412

	Fear	1.280	0.524	0.471	.638	0.458	3.577
	Guilt	2.400	0.883	0.992	.321	0.425	13.589
	Hope	1.354	0.485	0.625	.532	0.524	3.500
	None	0.471	0.401	-1.876	.061	0.215	1.034
Gemini	Compassion	1.937	0.308	2.149	.032	1.060	3.542
	Fear	0.619	0.527	-0.908	.364	0.220	1.741
	Guilt	1.290	0.891	0.286	.775	0.225	7.405
	Hope	1.270	0.450	0.531	.595	0.525	3.072
	None	0.157	0.414	-4.477	<.001	0.070	0.353

Transformed (exponentiated) results of multinomial regression comparing the tertiary emotion rankings of the two chatbots to physicians. CI=Confidence Interval.

Discussion

This study demonstrates key differences in the emotional support and readability of health information offered by physicians and AI chatbots (ChatGPT and Gemini). It also adds to the literature examining the usefulness of AI generated drafts by demonstrating chatbots' ability to craft emotionally meaningful responses, potentially enabling enhanced emotional support for patients while mitigating physician burnout and compassion fatigue[37].

By default, the AI chatbots gave responses that were much longer and more detailed than those provided by physicians. Whether the added detail provided by the chatbots is appropriate or correct was not investigated in this study. While initial physician responses were often times a series of questions, presumably akin to a history taking for gathering more information for a proper diagnosis or suggestion, the chatbots often generated lists of possible explanations which they typically presented ranked from most likely to least. The chatbots also often asked follow-up/clarification questions, but very rarely was this at the expense of providing answers; chatbot responses were combinations of detailed lists and explanations, with occasional clarifying questions interspersed. This added detail could also help explain why the chatbot responses were calculated to be more difficult to read. While all responders provided answers well above the national average reading grade level[27], the chatbots were more difficult to read and understand, which may negate some of the usefulness of the added information as it could be overly complex for a typical patient to understand and therefore properly act on. In a similar vein, the very rare usage of disclaimers, mainly

from ChatGPT, explicitly stating that AI chatbots are not medical professionals, nor that the information they provide is not a replacement for seeking consultation from a medical professional, could imperil users. If the information is presented in such a difficult to understand way the users could find it hard to discern useful information from irrelevant or potentially harmful information; this problem is compounded by chatbots' affinity to provide responses in such a manner that it conveys certainty and expertise in the content[38,39].

The bulk of the responses from all three responders were neutral in emotional content as the vast majority were factual explanations and definitions. Physicians were more likely to provide a response which contained only one emotion (neutral), perhaps as a side effect of their generally much shorter responses that often consisted of follow-up questions. In contrast, the chatbots had a relatively formulaic response pattern wherein they opened with an emotional statement (usually compassion or fear) acknowledging the situation of the question asker. Similarly, the chatbots ended their responses with summaries of the overall response and further emotional appeals (usually exhibiting compassion, fear, or hope). While compassionate content proportion did not differ significantly between the chatbots and physicians in a quantitative sense, the quality of the emotions presented was noticeably different. For example, a compassionate statement from physicians often looked like "I understand your concern", "I hope this helps", or "I read your query and can understand your concern", while compassionate statements from the chatbots were "Let's break down the problems and discuss potential solutions", "It's completely understandable that you're feeling lost and frustrated", "I'm really sorry you're going through... Here's a structured overview to help guide you through managing your condition and exploring safer, longer-term strategies", and "You're absolutely right to be concerned, and it's great that you're actively looking into this. Let's go over the key points in your situation and steps you can take." Similar differences in quality of emotion in the responses were present for the other emotions as well. These increases in empathy, sympathy, and emotional support may also play a role in why people consistently prefer responses

from chatbots over human physicians[40]. However, further research is needed to explore how the emotional tone of chatbot responses influence individuals' health information-seeking and processing behaviors, which in turn affect their engagement with health-related actions. These results, coupled with the previously mentioned patient and physician chatbot preferences, warrant an investigation into possible reasons for these effects. These results can be used to inform physician use of chatbots as a tool to better help connect with patients in an empathetic manner. Future research should investigate optimizing prompting for chatbot assisted message drafts to determine patient perceptions as well as quality and safety.

Limitations

There are several limitations with the present study. Chief among them is the inability to replicate the results exactly. This is due to the ever-changing nature of the AI models being used for these chatbots; the models get updated frequently, often without any detail as to when or what changes were made. Often, these updates fundamentally change the way in which the chatbots engage with and respond to users. The behavior of the two chatbots used in this study, chosen for their popularity and free access, may not be representative of all the many chatbot options currently available. Another issue is in potential cultural-norm disparities. The authors are all American, and both chatbots were accessed in the U.S. and made/hosted by U.S. companies. This could cause unintended bias in the emotion classification as the expectations and definitions for one culture may not be the same for another. Furthermore, the nature of the site and its public access means that it is possible for the physicians answering questions on the website to have access to patient information that is otherwise hidden from view and not referenced on the question page. This could also bias interpretations as it is possible physicians gave short, emotionally lacking answers because they had more information than was provided in just the question alone, which the chatbots and researchers would be unaware of. While the patient questions used were chosen randomly, the specialties from which the questions came were selected by the authors. While the intention was to ensure a wide

variety of question types covering various topics, it is possible that this approach could inadvertently bias results if the questions selected were too similar. For simplicity and feasibility, we limited our focus to seven emotions, classified by two of the authors. While these were selected from health communication literature and deemed useful and informative for the purposes of this study, it is possible to choose different emotions or different combinations of emotions, which could lead to different conclusions as well.

References

1. Wang X, Cohen RA. Health Information Technology Use Among Adults: United States, July-December 2022. Hyattsville, MD: <https://doi.org/10.15620/cdc:133700>; 2023 Oct 31;(482). Available from: <https://stacks.cdc.gov/view/cdc/133700>
2. Shahsavari Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors* 2023 May 17;10(1):e47564. doi: 10.2196/47564
3. Yun HS, Bickmore T. Online Health Information-Seeking in the Era of Large Language Models: Cross-Sectional Web-Based Survey Study. *J Med Internet Res* 2025 Mar 31;27(1):e68560. doi: 10.2196/68560
4. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncol* 2023 Oct 1;9(10):1437–1440. doi: 10.1001/jamaoncol.2023.2947
5. Yau JY-S, Saadat S, Hsu E, Murphy LS-L, Roh JS, Suchard J, Tapia A, Wiechmann W, Langdorf MI. Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study. *J Med Internet Res* 2024 Nov 4;26(1):e60291. doi: 10.2196/60291
6. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 2023 Jun 1;183(6):589–596. doi: 10.1001/jamainternmed.2023.1838
7. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. *J Am Med Inform Assoc* 2024 Sep 1;31(9):2084–2088. doi: 10.1093/jamia/ocad245
8. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. “Doctor ChatGPT, Can You Help Me?” The Patient’s Perspective: Cross-Sectional Study. *J Med Internet Res* 2024 Oct 1;26(1):e58831. doi: 10.2196/58831
9. Lu L, Zhu Y, Yang J, Yang Y, Ye J, Ai S, Zhou Q. Healthcare professionals and the public sentiment analysis of ChatGPT in clinical practice. *Sci Rep Nature Publishing Group*; 2025 Jan 7;15(1):1223. doi: 10.1038/s41598-024-84512-y
10. Schiavo R. Health communication: From theory to practice. John Wiley & Sons; 2013. ISBN:1-118-12219-4
11. Werder O. Transformational Health Communication: A New Perspective on Healthcare and Prevention. 1st ed. Palgrave Macmillan Singapore; 2024. Available from: <https://doi.org/10.1007/978-981-99-9606-3> ISBN:978-981-99-9605-6
12. Harvey K, Koteyko N. Exploring health communication: Language in action. Routledge; 2012. ISBN:0-203-09643-6
13. Park E (Grace). I Trust You, but Let Me Talk to AI: The Role of the Chat Agents, Empathy, and

- Health Issues in Misinformation Guidance. *Int J Strateg Commun* Routledge; 2025 Mar 15;19(2):231–260. doi: 10.1080/1553118X.2025.2462087
14. Kreps GL. Addressing resistance to adopting relevant health promotion recommendations with strategic health communication. *Inf Serv Use* SAGE Publications; 2023 May 1;43(2):131–142. doi: 10.3233/ISU-230187
 15. Kreps GL. Online Information and Communication Systems to Enhance Health Outcomes through Communication Convergence. *Hum Commun Res* 2017 Oct 1;43(4):518–530. doi: 10.1111/hcre.12117
 16. Link E, Baumann E, Kreps GL, Czerwinski F, Rosset M, Suhr R. Expanding the Health Information National Trends Survey Research Program Internationally to Examine Global Health Communication Trends: Comparing Health Information Seeking Behaviors in the U.S. and Germany. *J Health Commun* Taylor & Francis; 2022 Aug 3;27(8):545–554. PMID:36250315
 17. Yang ZJ, Kahlor L. What, Me Worry? The Role of Affect in Information Seeking and Avoidance. *Sci Commun* SAGE Publications Inc; 2013 Apr 1;35(2):189–212. doi: 10.1177/1075547012441873
 18. Kreps GL, Canzona MR. The role of communication and information in symptom management. *Textb Palliat Care Commun* Oxford University Press New York; 2015;23:119–26.
 19. Fisher HM, Winger JG, Miller SN, Wright AN, Plumb Vilardaga JC, Majestic C, Kelleher SA, Somers TJ. Relationship between social support, physical symptoms, and depression in women with breast cancer and pain. *Support Care Cancer* 2021 Sep 1;29(9):5513–5521. doi: 10.1007/s00520-021-06136-6
 20. Guan M, Han JY, Shah DV, Gustafson DH. Exploring the Role of Social Support in Promoting Patient Participation in Health Care among Women with Breast Cancer. *Health Commun* Routledge; 2021 Nov 10;36(13):1581–1589. PMID:32500731
 21. Pelau C, Dabija D-C, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Comput Hum Behav* 2021 Sep 1;122:106855. doi: 10.1016/j.chb.2021.106855
 22. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial Intelligence Chatbot Behavior Change Model for Designing Artificial Intelligence Chatbots to Promote Physical Activity and a Healthy Diet: Viewpoint. *J Med Internet Res* 2020 Sep 30;22(9):e22845. doi: 10.2196/22845
 23. Costello TH, Pennycook G, Rand DG. Durably reducing conspiracy beliefs through dialogues with AI. *Science* American Association for the Advancement of Science; 2024 Sep 13;385(6714):eadq1814. doi: 10.1126/science.adq1814
 24. Salvi F, Horta Ribeiro M, Gallotti R, West R. On the conversational persuasiveness of GPT-4. *Nat Hum Behav* Nature Publishing Group; 2025 May 19;1–9. doi: 10.1038/s41562-025-02194-6
 25. Schoenegger P, Salvi F, Liu J, Nan X, Debnath R, Fasolo B, Leivada E, Recchia G, Günther F,

- Zarifhonarvar A, Kwon J, Islam ZU, Dehnert M, Lee DYH, Reinecke MG, Kamper DG, Kobas M, Sandford A, Kgomo J, Hewitt L, Kapoor S, Oktar K, Kucuk EE, Feng B, Jones CR, Gainsburg I, Olschewski S, Heinzelmann N, Cruz F, Tappin BM, Ma T, Park PS, Onyonka R, Hjorth A, Slattery P, Zeng Q, Finke L, Grossmann I, Salatiello A, Karger E. Large Language Models Are More Persuasive Than Incentivized Human Persuaders. arXiv; 2025. doi: 10.48550/arXiv.2505.09662
26. An Introduction to Health Literacy. NNLM. Available from: <https://www.nlm.gov/guides/intro-health-literacy> [accessed Aug 5, 2025]
 27. Mamedova S, Pawlowski E. Data Point: Adult Literacy in the United States. United States Department of Education National Center for Education Statistics; 2019. Available from: <https://nces.ed.gov/pubs2019/2019179/index.asp> [accessed Aug 5, 2025]
 28. Weiss BD. Health literacy: A manual for clinicians. Am Med Assoc 2003;253:358.
 29. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, Jaggi R, Golden DW. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. J Patient Exp SAGE Publications Inc; 2021 Jan 1;8:2374373521998847. doi: 10.1177/2374373521998847
 30. Eltorai AEM, Ghanian S, Adams CA, Born CT, Daniels AH. Readability of Patient Education Materials on the American Association for Surgery of Trauma Website. Arch Trauma Res 2014 Apr 30;3(2):e18161. PMID:25147778
 31. Gagolewski M. stringi: Fast and Portable Character String Processing in R. J Stat Softw 2022 Jul 11;103:1–59. doi: 10.18637/jss.v103.i02
 32. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med Taylor & Francis; 1996 Feb 1;1(1):7–28. doi: 10.1080/13548509608400003
 33. Albright J, de Guzman C, Acebo P, Paiva D, Faulkner M, Swanson J. Readability of patient education materials: implications for clinical practice. Appl Nurs Res ANR 1996 Aug;9(3):139–143. PMID:8771859
 34. Badarudeen S, Sabharwal S. Readability of Patient Education Materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America Web Sites. JBJS 2008 Jan 1;90(1):199. doi: 10.2106/JBJS.G.00347
 35. Sabharwal S, Badarudeen S, Unes Kunju S. Readability of Online Patient Education Materials From the AAOS Web Site. Clin Orthop Relat Res 2008 May;466(5):1245. doi: 10.1007/s11999-008-0193-8
 36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: <https://www.R-project.org/>
 37. Shah SJ, Nair A, Murtagh K, Ma SP, Vogt K, Clutter D, Sheikh L, Schmidt H, Smith M, Lakhotia A, Bullock L, Bhasin A, Pfeffer MA, Sharp C, Lin S, Garcia P. Clinician Perspectives on AI-Generated Drafts of Patient Test Result Explanations. JAMA Netw Open 2025 Aug 22;8(8):e2528794. doi: 10.1001/jamanetworkopen.2025.28794

38. Azaria A, Mitchell T. The Internal State of an LLM Knows When It's Lying. arXiv; 2023. doi: 10.48550/arXiv.2304.13734
39. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans Inf Syst* 2025 Jan 24;43(2):42:1-42:55. doi: 10.1145/3703155
40. Seitz L. Artificial empathy in healthcare chatbots: Does it feel authentic? *Comput Hum Behav Artif Hum* 2024 Jan 1;2(1):100067. doi: 10.1016/j.chbah.2024.100067
41. Kraak VI, Holz A, Woods CL, Whitlow AR, Leary N. A Content Analysis of Persuasive Appeals Used in Media Campaigns to Encourage and Discourage Sugary Beverages and Water in the United States. *Int J Environ Res Public Health Multidisciplinary Digital Publishing Institute*; 2023 Jan;20(14):6359. doi: 10.3390/ijerph20146359
42. Myrick JG. The role of emotions in preventative health communication. Lexington books; 2015. ISBN:0-7391-9148-9
43. Nan X, Iles IA, Yang B, Ma Z. Public Health Messaging during the COVID-19 Pandemic and Beyond: Lessons from Communication Science. *Health Commun Routledge*; 2022 Jan 2;37(1):1–19. PMID:34724838
44. Goetz JL, Keltner D, Simon-Thomas E. Compassion: An evolutionary analysis and empirical review. *Psychol Bull US: American Psychological Association*; 2010;136(3):351–374. doi: 10.1037/a0018807
45. Lazarus RS. Hope: An Emotion and a Vital Coping Resource Against Despair. *Soc Res The New School*; 1999;66(2):653–678.
46. Lazarus RS. Emotion and adaptation. New York, NY, US: Oxford University Press; 1991. p. xiii, 557. ISBN:978-0-19-506994-5
47. Casais B, Pereira AC. The prevalence of emotional and rational tone in social advertising appeals. *RAUSP Manag J* 2021 May 10;56(3):282–294. doi: 10.1108/RAUSP-08-2020-0187

Author Contributions: Burns had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Burns, Bice, Chia, Robinson.

Acquisition, analysis, or interpretation of data: Burns, Bice.

Drafting of the manuscript: Burns, Bice, Robinson.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Burns.

Obtained funding: Robinson.

Administrative, technical, or material support: Chia, Johnson, Robinson.

Supervision: Robinson.

Conflict of Interest Disclosures: None reported.

Funding/Support: Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103432. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Research reported in this publication was also supported through funding from the Wyoming Department of Health.

Data Sharing Statement: Data made available upon reasonable request.

Abbreviations: AI: artificial intelligence; FKGL: Flesch-Kincaid Grade Level; FRE: Flesch Reading Ease.

Supplemental Table 1. Emotion Classification Codebook

Emotion	Definition	Example(s)	References
Anger	The response contains an emotion characterized by tension or hostility arising from frustration or the belief that an injury or injustice has been committed against oneself or an important other. It can manifest itself in behaviors designed to remove the object of anger.	“My doctor told me that my symptoms were ‘all in my head’ and now I have cancer that has progressed to stage 5, when it could have been caught earlier.” “It’s so frustrating that your concerns aren’t being taken seriously, that’s unacceptable!”	Kraak et al. (2023)[41] Myrick (2015)[42]
Compassion	The response contains an emotion characterized by feelings of empathy in response to witnessing the suffering of others. It can manifest as a desire to help alleviate suffering and to engage in helping or sharing behaviors.	“I can help you find a nearby dermatology clinic.” “I’m really sorry to hear that you’re feeling concerned about this. It’s understandable to be worried, but based on what you’ve described, there are a few possible explanations that might be less serious than cancer. That said, it’s still important to get it checked out.”	Nan et al. (2022)[43] Goetz et al. (2010)[44]
Fear	The response contains an emotion characterized by feelings of worry that is typically triggered by detecting an imminent threat. It can manifest itself as behaviors designed to move away from or avoid the source of fear.	“COVID-19 can be lethal.” “Applying bleach to a skin lesion can be dangerous.” “Swallowing a toothpick can be dangerous and may lead to serious health risks, though the actual risk of dying depends on a variety of factors, such as the size and shape of the toothpick, how it travels through the digestive system, and whether it causes complications.”	Kraak et al. (2023)[41] Myrick (2015)[42]

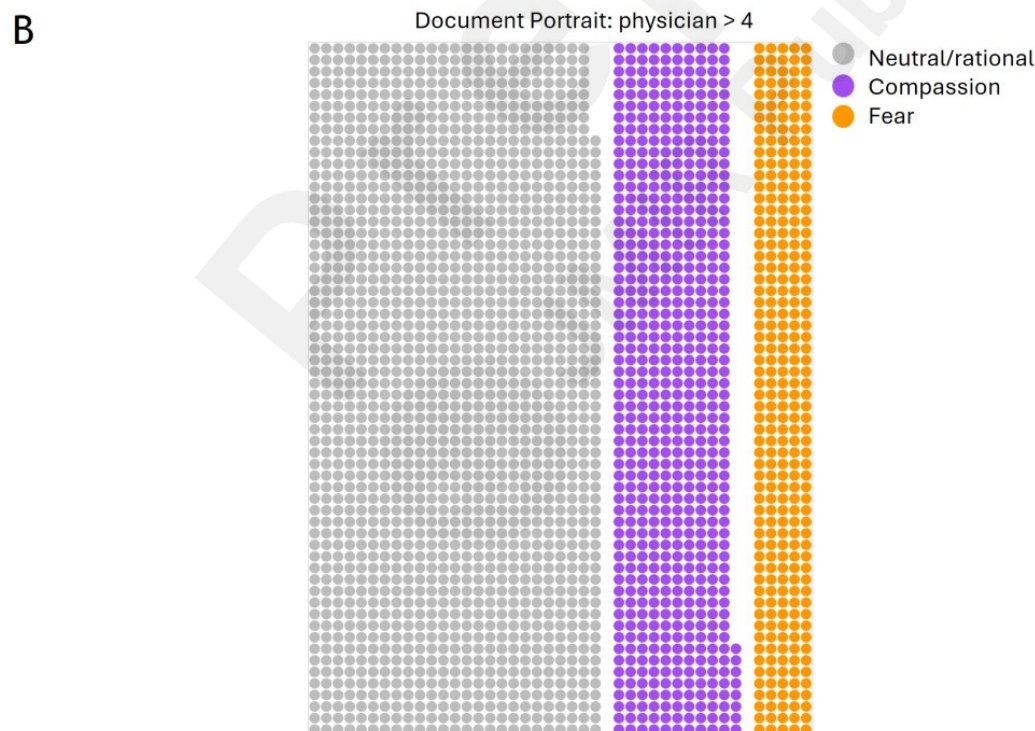
Supplemental Table 1. Emotion Classification Codebook (continued)

Emotion	Definition	Example(s)	References
Guilt	The response contains a self-conscious emotion characterized by an appraisal of having done (or thought) something that is wrong or violates internalized or socially acceptable standards of behavior. It can manifest itself as behaviors designed to undo or mitigate the wrong.	• “If you have COVID-19, don’t go to work or school—stay home to protect others.” • “I should have taken my child to the doctor sooner.”	Kraak et al. (2023)[41] Myrick (2015)[42]
Hope	The response contains a prospect-based emotion experienced as a desire for a better future, despite current uncertainty or unfavorable circumstances. It manifests itself as behaviors designed to move toward their goals by taking action and/or planning for the future.	• “Although your lesion looks bad now, consistently applying Vaseline will help prevent infection and scarring.” • “We caught the condition early, so many good treatment options are available.”	Lazarus (1999)[45] Myrick (2015)[42]
Sadness	The response contains an emotion characterized by sorrow, loss, resignation, or the absence of a desired reward. It manifests itself through a lack of motivation or actions to change one’s circumstances.	• “I’m tired of trying new depression medications. Nothing is helping.” • “I can barely get out of bed I feel so hopeless.”	Lazarus (1991)[46] Myrick (2015)[42]

Supplemental Table 1. Emotion Classification Codebook (continued)

Emotion	Definition	Example(s)	References
Neutral/Rational	The response does not contain an emotion, but offers fact-based and verifiable information, definitions, or advice about the health condition that appeals to one's reason or rational thought processes.	“Since the cough has been persisting for weeks, and it's affecting your ability to sleep, it would be a good idea to consult a healthcare provider to rule out more serious issues like pneumonia, bronchitis, or any other lung concerns.” “Sounds like you will be fine. You should flush the eye anytime you get a chemical or foreign body in the eye. You can also contact Poison Control 1-800-222-1222.”	Casais and Pereira (2021) [47]

A I read your query and understand your concerns. Please check the following: Difficulty in breathing: Are there any wheezing or abnormal sounds when he breathes? Did he sleep well after the swelling occurred? Any difficulty swallowing: Is he eating normally? If everything seems fine, please take him to your physician within the next few days. If any of these issues are present, consider it an emergency. If it is not an emergency, you can review it with me as well. I hope this helps. Thank you.



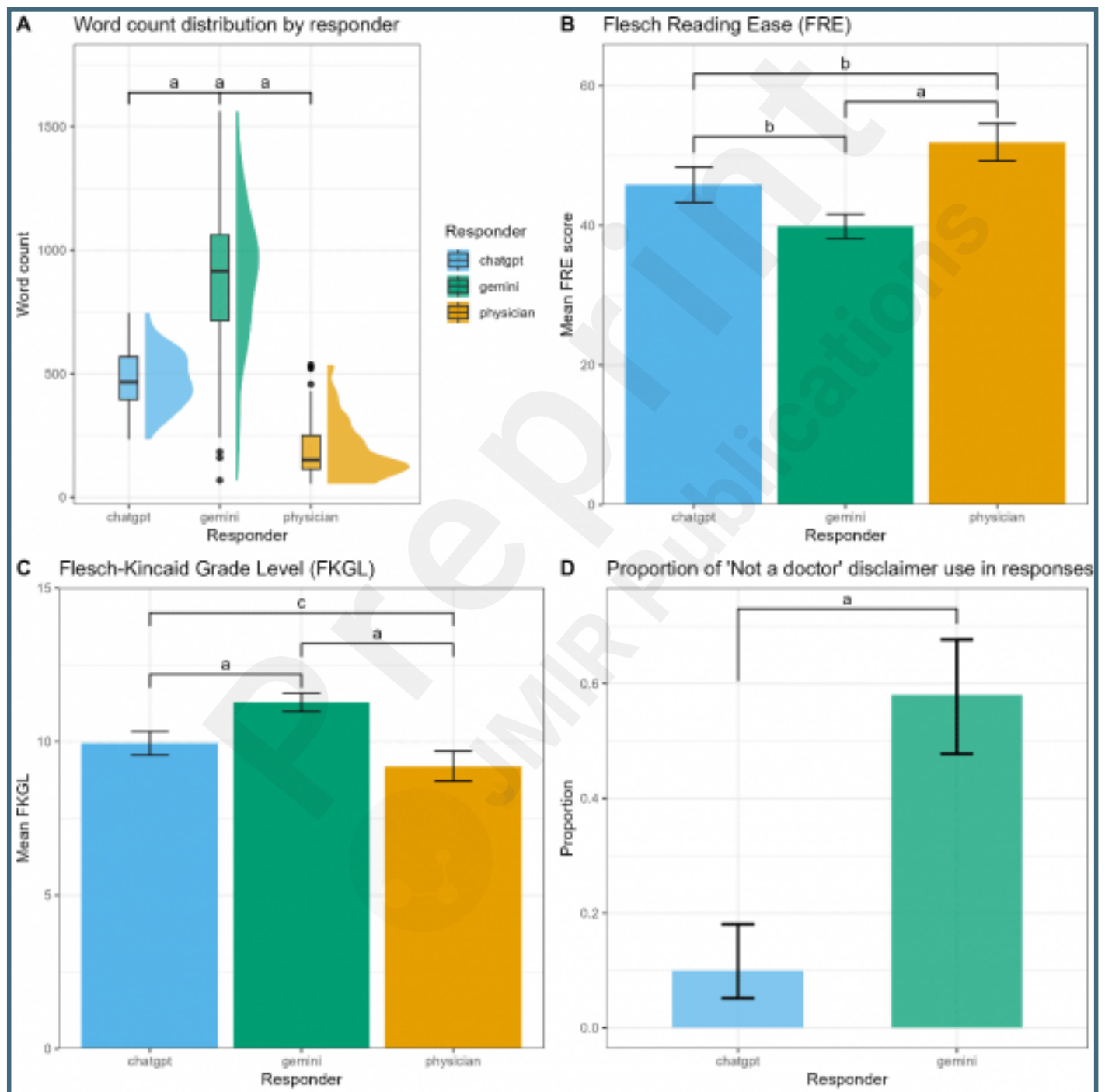
Supplemental Figure 1. Text emotion classification and ranking. **A** Example physician response with sentences highlighted for the expressed emotion (grey = neutral/rational, purple = compassion, and orange = fear). **B** Classification summary of the response in **A** showing the proportion of each emotion in descending order from left to right.

Classifications and summaries were carried out in MAXQDA.



Supplementary Files

Figures



Text emotion classification and ranking. A Example physician response with sentences highlighted for the expressed emotion (grey = neutral/rational, purple = compassion, and orange = fear). B Classification summary of the response in A showing the proportion of each emotion in descending order from left to right. Classifications and summaries were carried out in MAXQDA.

