Clinical Accuracy, Relevance, Clarity, and Emotional Sensitivity of Large Language Models to Surgical Patient Questions: Cross-Sectional Study

Mert Marcel Dagli, Felix Conrad Oettl, Jaskeerat Gujral, Kashish Malhotra, Yohannes Ghenbot, Jang W Yoon, Ali K Ozturk, William C Welch

Submitted to: JMIR Formative Research
on: January 08, 2024

Table of Contents

Clinical Accuracy, Relevance, Clarity, and Emotional Sensitivity of Large Language Models to Surgical Patient Questions: Cross-Sectional Study

Mert Marcel Dagli[1] MD; Felix Conrad Oettl[2, 3] MD; Jaskeerat Gujral[1]; Kashish Malhotra[4] MBBS; Yohannes Ghenbot[1] MD; Jang W Yoon[1] MD; Ali K Ozturk[1] MD; William C Welch[1] MD

[1]Department of Neurosurgery University of Pennsylvania Perelman School of Medicine Philadelphia US
[2]Department of Orthopedic Surgery Hospital for Special Surgery New York US
[3]Department of Orthopedic Surgery Schulthess Clinic Zurich CH
[4]Institute of Applied Health Research University of Birmingham Birmingham GB

Corresponding Author:
Mert Marcel Dagli MD
Department of Neurosurgery
University of Pennsylvania Perelman School of Medicine
801 Spruce Street
Philadelphia
US

Abstract

This cross-sectional study evaluates the clinical accuracy, relevance, clarity, and emotional sensitivity of responses to surgical patient inquiries provided by Large Language Models, highlighting their potential as adjunct tools in patient communication and education.

(JMIR Preprints 08/01/2024:56165)
DOI: https://doi.org/10.2196/preprints.56165

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 v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Title: Clinical Accuracy, Relevance, Clarity, and Emotional Sensitivity of Large Language Models to Surgical Patient Questions: Cross-Sectional Study

Authors: Dagli, Mert Marcel MD[1*]; Öttl, Felix MD[2,3]; Gujral, Jaskeerat[1]; Malhotra, Kashish MBBS[4]; Ghenbot, Yohannes MD[1]; Yoon, Jang W MD[1]; Ozturk, Ali K MD[1]; Welch, William C MD[1]

Author Affiliations:

1. Department of Neurosurgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, US
2. Department of Orthopedic Surgery, Hospital for Special Surgery, New York, NY, US
3. Department of Orthopedic Surgery, Schulthess Clinic, Zurich, CH
4. Institute of Applied Health Research, University of Birmingham, Birmingham, West Midlands, UK

Corresponding Author:

Mert Marcel Dagli, MD

Department of Neurosurgery

Perelman School of Medicine, University of Pennsylvania

801 Spruce Street, Philadelphia, 19107 PA, USA

Email: marcel.dagli@pennmedicine.upenn.edu

Phone: (267) 230-6493

Fax: (215) 615-3701

Keywords: artificial intelligence; natural language processing; large language model; generative AI; cross-sectional study; health information; patient education

Abstract word count: 73

Text word Count: 749

Number of references: 9

Number of tables and/or figures: 2

Number of videos: 0

Previous presentations: nil

Financial and competing interests disclosure: The authors declare no conflict of interests.

Funding: none

Abstract:

This cross-sectional study evaluates the clinical accuracy, relevance, clarity, and emotional sensitivity of responses to surgical patient inquiries provided by Large Language Models, highlighting their potential as adjunct tools in patient communication and education. Our findings demonstrated high performance of LLMs across accuracy, relevance, clarity, and emotional sensitivity, with Anthropic's Claude-2 outperforming OpenAI's ChatGPT and Google's Bard, suggesting LLMs' potential to serve as a complementary tool for enhanced information delivery and patient-surgeon interaction.

Introduction

Recent advances in natural language processing (NLP) have produced Large Language Model (LLM) applications, such as OpenAI's ChatGPT, that have captivated a worldwide audience [1]. These advancements have permeated the healthcare sector offering several benefits [2]. While LLMs have immense potential in improving clinical practice and patient outcomes, their role has not been completely established [3]. Often, patients that require surgery struggle with complex, anxiety-inducing questions [4]. Thus, preoperative counseling during preoperative workup is of utmost importance for informed consent, establishing trust, and pre-surgical optimization to improve patient outcomes. This process, being resource-intensive and involving numerous conversations, often leads to delays in communication that can be a significant source of frustration for patients [5]. Therefore, the importance of clear, adequate, and timely information delivery cannot be overemphasized. LLMs with chat features could improve preoperative communication, however, LLMs' ability in answering patients' surgical questions have not been extensively studied yet. Thus, this study aims to assess LLMs' potential and proficiency in responding to surgical patient questions.

Methods

In the formulation of our questionnaire, we utilized the input of three neurosurgical attendings, focusing on common general patient inquiries regarding surgery. 38 patient questions were presented in web sessions to three publicly accessible LLMs, OpenAI's ChatGPT GPT-4, Anthropic's Claude 2, and Google's Bard on August 16, 2023 (Multimedia Appendix 1). Questions revolved around four central themes: understanding the nature and rationale of surgery, pre-operative concerns, procedural aspects, and post-operative considerations. Each reply from the LLMs was reviewed by two independent blinded reviewers (MMD, FCO; research fellows who have medical doctorates but have not completed post-graduate clinical trianing). A 5-point Likert scale was used to assess accuracy, relevance, and clarity of responses [6]. Emotional sensitivity was evaluated on a 7-point Likert scale to increase discriminatory power [7]. Assessment of data normality was conducted using the Shapiro-Wilk test. Homogeneity of variances (homoscedasticity) across groups was evaluated via the Levene test. For non-parametric analysis, the Kruskal-Wallis test was employed to discern differences among groups. Subsequent pairwise comparisons were facilitated by the post-hoc Dunn test. In instances where parametric assumptions were upheld, a one-way ANOVA was conducted, followed by post-hoc analysis with Tukey's Honestly Significant Difference (HSD) test. *P* values of post-hoc analysis were adjusted for multiplicity with Bonferroni correction. Additionally, Weighted Percentage Agreement (WPA) was calculated to provide information on agreement levels between raters. All statistical analysis was performed using Python, version 3.7 (Python Foundation).

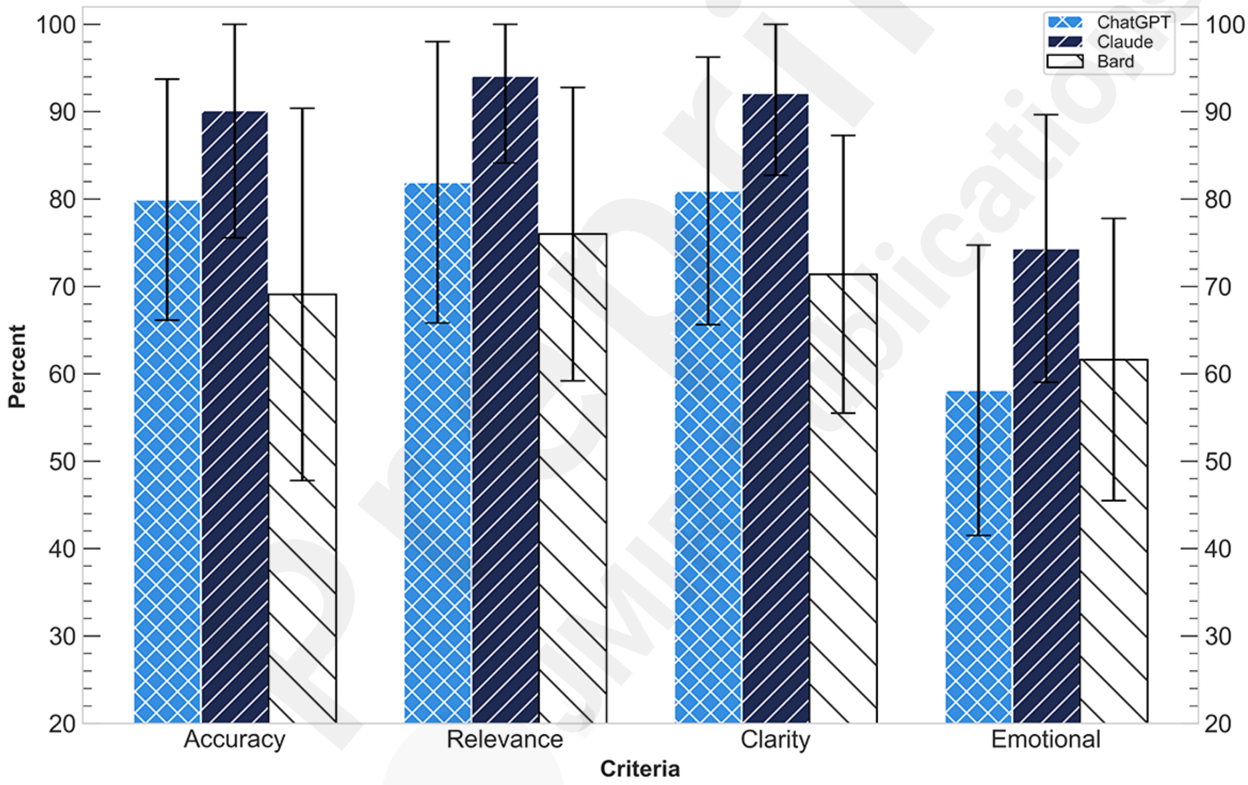
Ethical considerations

The study qualified for institutional review board (IRB) exemption as it exclusively utilized questions sourced from surgeon input, with no direct patient involvement.

Results

Shapiro-Wilk testing indicated non-normality (P<.05; Table 1) for accuracy, relevance, and clarity scores. Levene testing revealed non-homoscedasticity for relevance (F_2=5.009; P=.008). Kruskal-Wallis test showed significant differences in the distribution of accuracy (H=27.464; P<.001), relevance (H=29.074; P<.001), and clarity (H=32.745; P<.001). Post hoc Dunn test demonstrated that Claude's responses were significantly higher rated than those from ChatGPT and Bard in accuracy, relevance, and clarity (P<.05). There were no significant differences between ChatGPT and Bard, except for the clarity criterion (Z=1.972, P=.038). ANOVA showed significant differences in emotional sensitivity (F=10.799; P<.001). Post-hoc Tukey's HSD revealed significantly higher emotional sensitivity scores for Claude compared to ChatGPT and Bard (P<.05). WPA was highest for Claude followed by ChatGPT and Bard (Figure 1).

Figure 1: Bar chart of adjusted percentage average ratings of large language model responses (ChatGPT=light blue; Claude=dark blue; Bard=white). All mean Likert scale and adjusted percentage ratings (%) with their standard deviations are shown in the first table in the lower section of the figure. Adjusted average percentage ratings were calculated as the mean of normalized scores, using the formula: Adjusted Percentage Rating = ((Actual Likert Score - 1) / (Likert Scale Maximum - 1)) x 100%, to scale responses uniformly from 0 to 100%. The second table includes the weighted percentage agreement (WPA) point estimates with their 95% confidence intervals.



	ChatGPT Likert	ChatGPT %	Claude Likert	Claude %	Bard Likert	Bard %
Accuracy	4.2 (0.55)	79.93 (13.8)	4.61 (0.58)	90.13 (14.58)	3.76 (0.85)	69.08 (21.3)
Relevance	4.28 (0.64)	81.91 (16.1)	4.76 (0.4)	94.08 (9.96)	4.04 (0.67)	75.99 (16.79)
Clarity	4.24 (0.61)	80.92 (15.31)	4.68 (0.38)	92.11 (9.38)	3.86 (0.64)	71.38 (15.89)
Emotional	4.49 (1)	58.11 (16.61)	5.46 (0.92)	74.34 (15.3)	4.7 (0.97)	61.62 (16.16)

	ChatGPT WPA	Claude WPA	Bard WPA
Accuracy	80.26 (67.61-92.92)	86.84 (76.09-97.59)	71.05 (56.63-85.47)
Relevance	76.32 (62.8-89.83)	97.37 (92.28-102.46)	71.05 (56.63-85.47)
Clarity	72.37 (58.15-86.59)	94.74 (87.64-101.84)	60.53 (44.98-76.07)
Emotional	68.42 (53.64-83.2)	77.63 (64.38-90.88)	67.11 (52.17-82.04)

Table 1. Results of Normality Test (Shapiro-Wilk), Homoscedasticity Test (Levene), Nonparametric Test (Kruskal-Wallis), Post Hoc Pairwise Comparison of Nonparametric Data (Dunn Test with Bonferroni Correction), Parametric Test (Analysis of Variance), and Post Hoc Pairwise Comparison of Parametric Data (Tukey's Honestly Significant Differences Test with Bonferroni Correction).

Test	Value	*P* value
Shapiro-Wilk		
ChatGPT Accuracy, *W* statistic	0.862	<.001
Claude Accuracy, *W* statistic	0.711	<.001
Bard Accuracy, *W* statistic	0.87	<.001
ChatGPT Relevance, *W* statistic	0.845	<.001
Claude Relevance, *W* statistic	0.604	<.001
Bard Relevance, *W* statistic	0.917	.008
ChatGPT Clarity, *W* statistic	0.886	.001
Claude Clarity, *W* statistic	0.747	<.001
Bard Clarity, *W* statistic	0.933	.024
ChatGPT Emotional sensitivity, *W* statistic	0.965	.27
Claude Emotional sensitivity, *W* statistic	0.953	.11
Bard Emotional sensitivity, *W* statistic	0.959	.181
Levene		
Accuracy, F_2 statistic	2.144	.122
Relevance, F_2 statistic	5.009	.008
Clarity, F_2 statistic	1.918	.152
Emotional sensitivity, F_2 statistic	0.184	.833
Kruskal-Wallis		
Accuracy, *H* statistic	27.363	<.001
Relevance, *H* statistic	29.074	<.001
Clarity, *H* statistic	32.745	<.001
Dunn Test with Bonferroni		
Accuracy, ChatGPT vs Claude, *Z* statistic	-2.546	.004
Accuracy, ChatGPT vs Bard, *Z* statistic	1.56	.147
Accuracy, Claude vs Bard, *Z* statistic	4.106	<.001
Relevance, ChatGPT vs Claude, *Z* statistic	-2.872	<.001
Relevance, ChatGPT vs Bard, *Z* statistic	1.235	.342
Relevance, Claude vs Bard, *Z* statistic	4.107	<.001
Clarity, ChatGPT vs Claude, *Z* statistic	-2.546	.004
Clarity, ChatGPT vs Bard, *Z* statistic	1.972	.038
Clarity, Claude vs Bard, *Z* statistic	4.518	<.001
Analysis of Variance (ANOVA)		

Emotional sensitivity, *F* statistic	10.799	<.001
Tukey's HSD Test with Bonferroni		
Emotional sensitivity, ChatGPT vs Claude, *Q* statistic	-0.974	<.001
Emotional sensitivity, Bard vs ChatGPT, *Q* statistic	0.21	.607
Emotional sensitivity, Claude vs Bard, *Q* statistic	0.763	.002

Discussion

Our investigation revealed a promising potential for the use of LLMs for patient education. Anthropic's Claude-2 had significantly higher percentage average ratings of above 90% for accuracy (*P*=.004, *P*<.001), relevance (*P*<.001), and clarity (*P*=.004, *P*<.001), compared to ChatGPT and Bard. It also scored significantly better on emotional sensitivity than ChatGPT and Bard (*P*<.001, *P*=.002), with 74.3%. In a study parallel to ours, Sezgin et al. assessed the clinical accuracy of LLMs in the context of postpartum depression, demonstrating their efficacy in providing clinically accurate information, a finding that complements our study's illustration of LLMs' potential in patient education and engagement [8]. By providing accurate and timely information, LLMs can potentially alleviate patient concerns.

Limitations

The study's limitations include the absence of direct patient input in questionnaire formulation, lack of repeated zero-shot questioning which may reveal variability, and no dedicated analysis of overtly inaccurate hallucinations. The principal challenge for LLM deployment in clinical settings lies in its regulatory approval and secure integration within healthcare systems [9]. We are actively conceptualizing a randomized clinical trial (RCT), controlling for these limitations, to investigate LLM and surgeon responses as rated by patients and surgeons.

Conclusions

While surgeons remain indispensable in patient education, LLMs can potentially serve as a complementary tool, enhancing information delivery and supporting patient-surgeon interactions.

Authors' Contributions

WCW is the guarantor of the study. MMD and WCW led conceptualization, data acquisition, analysis, drafting and revision of the manuscript. JG and KM contributed to data acquisition, analysis, and drafting. Blinded scoring was performed by MMD and FCO. All authors contributed to analysis, interpretation, and drafting. JWY, AKO, and WCW contributed critical guidance at all stages of the study. The manuscript was reviewed, edited, and its final version approved by all authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Data Availability

All data generated or analyzed during this study are included in this published article (Multimedia Appendix 1).

Abbreviations

ANOVA: analysis of variance

EMR: electronic medical record

HSD: honestly significant difference

IRB: institutional review board

LLM: large language model

NLP: natural language processing

RCT: randomized clinical trial

WPA: weighted percentage agreement

References

1. Kevin R. The Brilliance and Weirdness of ChatGPT. 2022 [cited 2023 August 28]; Available from: https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html.

2. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019 Jun;6(2):94-8. PMID: 31363513. doi: 10.7861/futurehosp.6-2-94.

3. Mofatteh M. Neurosurgery and artificial intelligence. AIMS Neurosci. 2021;8(4):477-95. PMID: 34877400. doi: 10.3934/Neuroscience.2021025.

4. Wongkietkachorn A, Wongkietkachorn N, Rhunsiri P. Preoperative Needs-Based Education to Reduce Anxiety, Increase Satisfaction, and Decrease Time Spent in Day Surgery: A Randomized Controlled Trial. World J Surg. 2018 Mar;42(3):666-74. PMID: 28875242. doi: 10.1007/s00268-017-4207-0.

5. Williams S, Weinman J, Dale J. Doctor-patient communication and patient satisfaction: a review. Fam Pract. 1998 Oct;15(5):480-92. PMID: 9848436. doi: 10.1093/fampra/15.5.480.

6. Sullivan GM, Artino AR, Jr. Analyzing and interpreting data from likert-type scales. J Grad Med Educ. 2013 Dec;5(4):541-2. PMID: 24454995. doi: 10.4300/JGME-5-4-18.

7. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychol (Amst). 2000 Mar;104(1):1-15. PMID: 10769936. doi: 10.1016/s0001-6918(99)00050-5.

8. Sezgin E, Chekeni F, Lee J, Keim S. Clinical Accuracy of Large Language Models and Google Search Responses to Postpartum Depression Questions: Cross-Sectional Study. J Med Internet Res. 2023 Sep 11;25:e49240. PMID: 37695668. doi: 10.2196/49240.

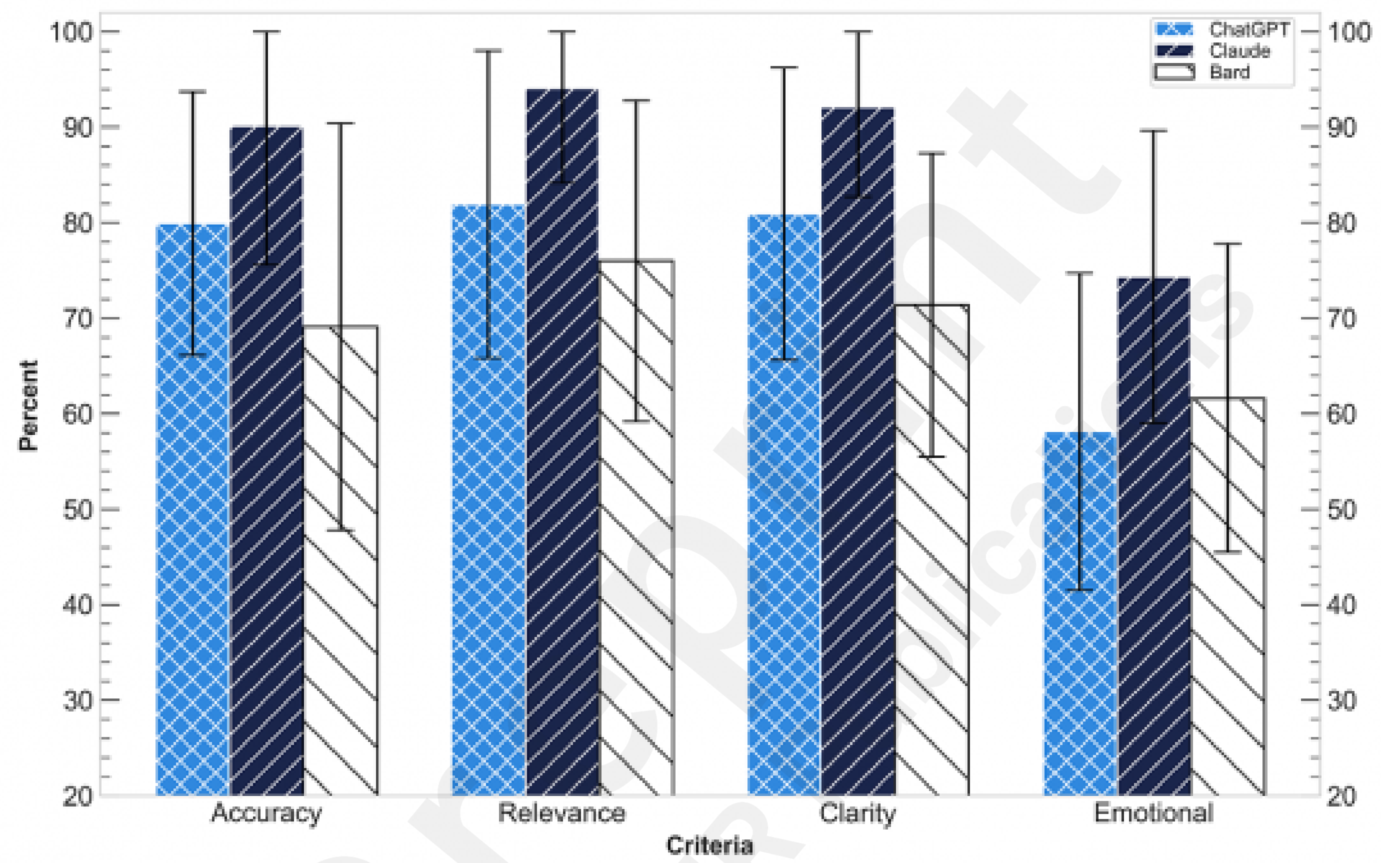
9. Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. Journal of family medicine and primary care. 2019;8(7):2328. PMID: 31463251. doi:

10.4103/jfmpc.jfmpc_440_19.

Supplementary Files

Figures

Bar chart of adjusted percentage average ratings of large language model responses (ChatGPT=light blue; Claude=dark blue; Bard=white). All mean Likert scale and adjusted percentage ratings (%) with their standard deviations are shown in the first table in the lower section of the figure. Adjusted average percentage ratings were calculated as the mean of normalized scores, using the formula: Adjusted Percentage Rating = ((Actual Likert Score - 1) / (Likert Scale Maximum - 1)) x 100%, to scale responses uniformly from 0 to 100%. The second table includes the weighted percentage agreement (WPA) point estimates with their 95% confidence intervals.



	ChatGPT Likert	ChatGPT %	Claude Likert	Claude %	Bard Likert	Bard %
Accuracy	4.2 (0.55)	79.93 (13.8)	4.61 (0.58)	90.13 (14.58)	3.76 (0.85)	69.08 (21.3)
Relevance	4.28 (0.64)	81.91 (16.1)	4.76 (0.4)	94.08 (9.96)	4.04 (0.67)	75.99 (16.79)
Clarity	4.24 (0.61)	80.92 (15.31)	4.68 (0.38)	92.11 (9.38)	3.86 (0.64)	71.38 (15.89)
Emotional	4.49 (1)	58.11 (16.61)	5.46 (0.92)	74.34 (15.3)	4.7 (0.97)	61.62 (16.16)

	ChatGPT WPA	Claude WPA	Bard WPA
Accuracy	80.26 (67.61-92.92)	86.84 (76.09-97.59)	71.05 (56.63-85.47)
Relevance	76.32 (62.8-89.83)	97.37 (92.28-102.46)	71.05 (56.63-85.47)
Clarity	72.37 (58.15-86.59)	94.74 (87.64-101.84)	60.53 (44.98-76.07)
Emotional	68.42 (53.64-83.2)	77.63 (64.38-90.88)	67.11 (52.17-82.04)

Multimedia Appendixes

Average ratings of large language model responses for accuracy, relevance, clarity, and emotional sensitivity.
URL: http://asset.jmir.pub/assets/05d55516653e0a706a8d94997492d913.xlsx

Responses to surgical patient questions.
URL: http://asset.jmir.pub/assets/ba46415621f057f797434a4d554e863d.xlsx