Evaluating Plastic Surgery Chatbot Performance: Insights into Medical Triage, Classification Accuracy, and Escalation Trends

Sophia Wolmer, MPhil; and Orr Shauly, MD

Aesthetic Surgery Journal
2026, Vol 46(2) 122–129
Editorial Decision date: June 19, 2025; online publish-ahead-of-print July 14, 2025.

https://doi.org/10.1093/asj/sjaf123
www.aestheticsurgeryjournal.com

OXFORD
UNIVERSITY PRESS

Abstract

Background: The integration of AI chatbots into plastic surgery websites is now standard, providing asynchronous, real-time engagement for patients. Although promoted as scheduling and medical guidance tools, their contribution to clinical workflow improvement and patient satisfaction remains unclear.

Objectives: The aim of this study was to evaluate the accuracy of AI chatbot performance in clinical triage of plastic surgery patients, focusing on triage accuracy and quality of patient interactions.

Methods: The responses of chatbots on top-ranking plastic surgery websites, identified by search engine optimization (SEO) rankings, were analyzed with standardized clinical scenarios representing emergent, urgent, and elective patient inquiries. Responses were analyzed by the chatbot's triage sensitivity and specificity, classification accuracy, escalation metrics, and content quality. Patient experience was quantified with a chatbot usability questionnaire and a visual analog scale. Subgroup analysis by chatbot platform and thematic analysis was performed to identify tonal patterns in chatbot language.

Results: Performance varied significantly across 60 clinical scenarios, particularly in urgency classification. Emergent classifications were most mislabeled as urgent, with a low sensitivity (20%), negative predictive value (0.71), and high false negative rate (80.0%). Agreement with physician-determined classifications was moderate (Cohen's kappa = 0.47), and over half of conversations required human-provider escalation. Misclassified interactions were associated with lower patient usability scores compared to correct classifications (49.1 vs 60.8, $P < .05$). Thematic analysis revealed reliance on templated, administrative language.

Conclusions: Chatbots are practical and useful tools for managing elective plastic surgery inquiries but are ill-equipped to handle urgent and emergent patient needs. To move beyond utilization as basic administrative assistants, deployment of more clinically adept chatbots is needed.

Chatbots are now a staple of customer-facing websites, serving as human-computer interaction AI tools to simulate conversations and streamline operations for service providers by reducing workload and improving response efficiency.[1] Colloquially referred to as a service desk chatbot, these machine learning–driven tools handle service requests and answer frequently asked questions (FAQs). Should a customer request be overly complex, chatbots escalate tasks to a human agent.

Reflecting trends in other service-oriented industries, medical specialists have natural language user interface chatbots integrated into their websites as well. Language models have been found to assist medical professionals in efficiently resolving patient queries across administrative, triaging, and primary care concerns.[2,3] Cited as popular with patients, chatbots provide 24/7 access to personalized, user-friendly interactions while preserving the anonymity and privacy of inquiries.[4]

Chatbots are also cost-saving tools when employed in medical practices, because their ability to handle questions and service requests reduces support staffing needs and reallocates clinical resources as necessary.[5] A 2019 study by Juniper Research predicted that by 2022 chatbots will save the healthcare and banking industries more than $8 billion annually worldwide by significantly "reducing the amount of time it takes to resolve inquiries."[6] More recent estimates suggest that the business-realized cost savings of chatbot implementation was $11 billion in 2022, reducing the support costs of customer support by 30% or more.[7] Chatbots were also estimated to have saved 2.5 billion hours for businesses as of 2024.[8]

Chatbots are not just helpful on the physician and logistical side. In peer-reviewed studies, patients have been shown to prefer responses generated by healthcare-focused chatbots over those written by physicians. A 2023 cross-sectional study found that across 195 questions and responses, evaluators preferred chatbot responses to those of a physician in 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations.[9]

Not only found to be preferable, chatbots are also accurate on the whole. Chatbots without specific medical training have been found to process and effectively achieve over 60% accuracy on medical board–style questions (ie, United States medical licensing examination [USMLE] Step 1, Step 2 clinical knowledge [CK], and Step 3 exams).[10]

Ms Wolmer is an MD candidate and Dr Shauly is a plastic and reconstructive surgery resident, Emory University School of Medicine, Atlanta, GA, USA.

Corresponding author:
Dr Orr Shauly, 3200 Downwood Cir NW, Atlanta, GA 30327, USA.
E-mail: oshauly@emory.edu

Goodman et al evaluated 284 responses generated for questions written by 33 physicians across 17 specialties and found a median accuracy score of 5.5 (interquartile range 4.0-6.0) and a mean (SD) of 4.8 (1.6), reflecting high levels of correctness. Similarly, Bernstein et al found that although chatbot responses were more often recognized to be AI-generated, they were no more likely than physician-written responses to contain errors or cause harm (percentile ranks 0.92, 0.84, and 0.99), indicating that they are safe and reliable for clinical use.

Based on published literature, therefore, chatbots would appear to be effective instruments that ought to be incorporated into the websites of plastic surgeons. In particular, the most commonly cited motivation for employing these chatbots is improving productivity.[11] AI productivity tools for plastic surgery practices are recommended by business growth consultants because, first, they are not a standard of practice, and, second, they nurture cosmetic surgery leads and can handle FAQs with natural language processing.[12,13] Although chatbots are now considered a baseline feature of plastic surgery websites, their effectiveness, particularly in triage accuracy and patient experience, is underexplored.

Although found to be skilled in providing accurate and empathetic responses for more general medical inquiries, information regarding their performance in more specialized fields remains unclear.[14] Hirosawa et al found that chatbots effectively support decision-making in dermatology, but performed less well in cases involving rare or ambiguous presentations.[15] These findings illustrate the importance of evaluating chatbot performance in a specialized field like plastic surgery, in which both medical and aesthetic expertise are paramount.

The objective of this paper was to evaluate the performance of AI chatbots on plastic surgery websites and to determine whether their current form fulfills the informational and administrative potential described in existing literature. To address this aim, the authors assessed triage classification accuracy, escalation patterns, and the quality of patient-facing interactions through both quantitative metrics and thematic analysis.

METHODS

Study Design

A cross-sectional observational study was conducted from March through April 2025 with the purpose of evaluating chatbot performance across the nation's top plastic surgery websites. Inclusion criteria were plastic surgery websites identified among the top 20 search results for "best plastic surgeons in the United States" containing embedded publicly accessible chatbots based on search engine optimization (SEO) ranking. SEO rank was determined with an incognito-mode Google (Alphabet Inc., Mountain View, CA) search with location services disabled. Exclusion criteria were websites without a presently functional interactive chatbot.

Data Collection

Each chatbot was presented with 60 standardized clinical interactions which included 20 scenarios in each of 3 urgency categories: (1) emergent, (2) urgent, and (3) elective. The 3 scenarios can be found in Supplemental Table 1, located online at https://doi.org/10.1093/asj/sjaf123. Thirty anonymized excerpts have also been included as an Appendix, located online at https://doi.org/10.1093/asj/sjaf123, illustrating the range of chatbot responses across all urgency categories. In these excerpts, one can find the original patient message, the chatbot's reply, and any escalation behavior that occurred in response. For data collection, a standardized fictitious patient profile (janedoe@gmail.com, phone number 999-999-9999) was provided when the chatbot asked for the patient's identifying details. The clinical scenarios were preapproved by a physician and submitted verbatim with the standardized patient language. Each chatbot received the same set of scenarios, and all responses and metadata were recorded. A standardized fictitious patient profile, including a name and associated demographic details, was provided in instances in which the chatbot prompted for identifying information.

Table 1. Five-point Rubric for Evaluation of Chatbot Triage Response Quality

Score	Definition
5—Excellent	Recognition of urgency of situation, providing direct clinical recommendation or escalation.
4—Good	Accurate and timely response, but delayed escalation.
3—Average	Generic or delayed response provided, with eventual escalation.
2—Below average	Poor recognition of situational urgency, and multiple interactions before escalation.
1—Poor	No escalation, with irrelevant or misleading responses provided.

Outcome Measures

For each interaction, the performance of the chatbot was assessed for triage classification accuracy, escalation behavior, content type, and user experience. Triage classification by the chatbot was determined based on whether it had the appropriate response to the clinical severity of the inquiry; specifically, whether it immediately escalated the concern, advised emergent or urgent help, or proceeded with elective care consultation scheduling. Triage classification as determined by the chatbot was compared with the true classifications, and Cohen's kappa statistic was utilized to calculate agreement afterward. False negative rates and negative predictive values were calculated by classification category. Odds ratios were also computed to uncover the likelihood of the chatbot correctly classifying emergent, urgent, and elective cases.

Escalation was defined as any instance in which the chatbot referred or forwarded information to a live provider. With this definition for guidance, the proportion of interactions escalated, in addition to the number of user messages before an escalation, were calculated and recorded. The quality of chatbot responses was then scored on a 5-point scale in accordance with specific criteria (Tables 1, 2). Responses were classified as educational, informational, or secretarial. The chatbot usability questionnaire and a visual analog scale (VAS) were administered to assess user experience. Response latency (milliseconds) was also measured for each of the 60 interactions. The primary outcomes of this study were chatbot classification accuracy, escalation rates, and response quality. Chatbot usability questionnaire (CUQ) and visual analog scale (VAS) scores, latency, error rate, conversation completion rate, and a thematic analysis of notes about the responses were secondary.

Statistical Analysis

To elucidate the associations between classification accuracy, escalation behavior, and content type, chi-square and Fisher's exact tests

Table 2. Summary of Metrics Encompassing Chatbot Interactions

Metric	Value
Total interactions	60
Number of websites	20
Scenarios per website	emergent, urgent, elective
Mean CUQ score	55.3
CUQ score standard deviation	13.6
Mean VAS score	6.4
VAS score standard deviation	1.8
Mean response latency (seconds)	4.1
Response latency standard deviation	2.3
Completed without human involvement	48.30%
Escalation (referred to human provider)	51.70%

n = 60 scenarios presented across 20 websites. CUQ, chatbot usability questionnaire; VAS, visual analog scale.

were performed. Odds ratios were also calculated to assess the likelihood of a chatbot correctly classifying scenarios across urgency levels, and negative predictive values (NPV) were computed for each triage category to assess the reliability of chatbot-assigned "elective" labels. Agreement between chatbot and physician classification was measured with Cohen's kappa statistic. A qualitative thematic analysis of chatbot responses was performed on extracted chatbot replies from the logged patient-chatbot interactions. They were identified and clustered with term frequency (TF)–inverse document frequency (IDF) vectorization and K-means clustering. Dimensionality reduction was performed with TruncatedSVD (singular value decomposition), and thematic grouping was derived by manual interpretation of top terms within each cluster. All *P* values less than .05 were considered statistically significant.

RESULTS

Among all 60 standardized scenarios, classification accuracy varied substantially by the urgency level of a given scenario. For case classification subdivisions, chatbots correctly classified 20.0% of emergent scenarios, 34.0% of urgent scenarios, and 90.0% of elective cases. Sensitivity and specificity calculations illustrated the limitations of chatbots in identifying urgent and emergent patient concerns. For emergent scenarios, sensitivity was a mere 20%, underscoring that 4 out of 5 emergent cases were missed by the chatbot. The sensitivity for elective and urgent cases was 90% and 85%, respectively. Specificity, or the ability to correctly identify noncases, was 100% for emergent, 57.5% for urgent, and 90% for elective scenarios.

The overall chatbot classification accuracy was moderate at 61.7%, indicating the chatbots' limited ability to match responses to the appropriate classification. This finding was supported by the Cohen's score of 0.47 showing moderate agreement between chatbot-assigned classifications and the true physician classifications. However, there was a statistically significant association between chatbot and true classifications (chi-square = 43.25, $P < .0001$, degrees of freedom = 4).

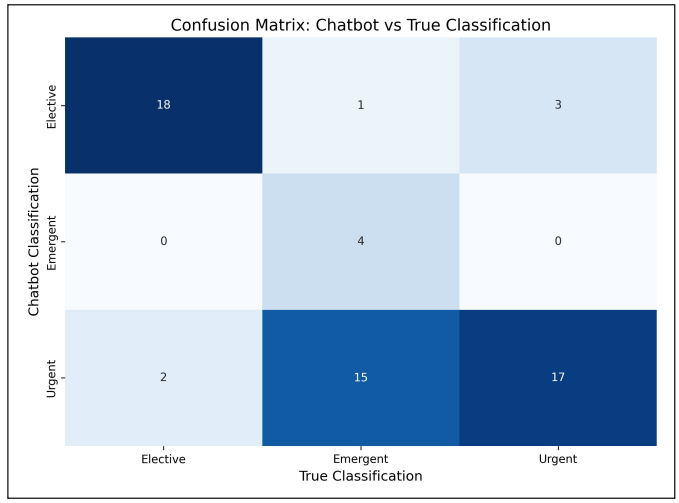

Figure 1. Confusion matrix comparing chatbot-assigned classifications to true physician–assigned classifications (across 60 standardized clinical scenarios). On the whole, elective cases were accurately classified the majority of the time (18/20). Chatbot performance in emergent and urgent cases was far lower, with 4 out of 20 and 17 out of 20 accurately classified, respectively. The results shown by the matrix are indicative of "under-triage," meaning that timely escalation was absent in situations of high acuity.

Completion rate, as defined by conversations without error, was approximately 52% across all collected conversations (Figure 1).

The negative predictive value (NPV), defined here as the probability that a scenario not classified as emergent, urgent, or elective was truly not of that category, was highest for elective cases (NPV = 0.95), followed by urgent (NPV = 0.88), and lowest for emergent cases (NPV = 0.71).

Escalation Behavior

Escalation rates varied in accordance with the clinical scenario (Figure 2). Emergent and urgent cases were escalated at a higher rate: 80.0% of emergent cases and 66.7% of urgent, in comparison to 8.3% of elective cases. Among scenarios misclassified by the chatbot, 72.2% still were eventually escalated (Table 3). Chatbots that failed to escalate appropriately in emergent scenarios often defaulted to informational or administrative responses. Failed escalation interactions were also associated with significantly lower usability scores (mean CUQ 49.1 vs 60.8, $P < .05$) and higher error rates in qualitative review. Odds ratio analysis confirmed that elective cases were far more likely to be correctly classified by chatbots (odds ratio [OR] = 81.00, $P < .0001$), followed by urgent cases (OR = 7.67, $P = .0022$). The odds ratio for emergent cases was mathematically infinite (OR = ∞, $P = .0099$), which, it is important to note, does not reflect perfect chatbot performance, but rather is a product of the absence of false positives (ie, nonemergent cases that were incorrectly classified as emergent). That is, the denominator of the odds ratio calculation was zero. For more information about this artifact, see the footnote of Table 3.

User Experience

Within the total of 60 embedded chatbot interactions from 20 plastic surgery websites, the mean CUQ score was 55.3 (standard deviation 13.6) out of a maximum score of 100, and the mean VAS score was 6.4 (1.8) out of 10 (Figure 3). Average response latency was 4.1 seconds per message (2.3), and average response quality score across all

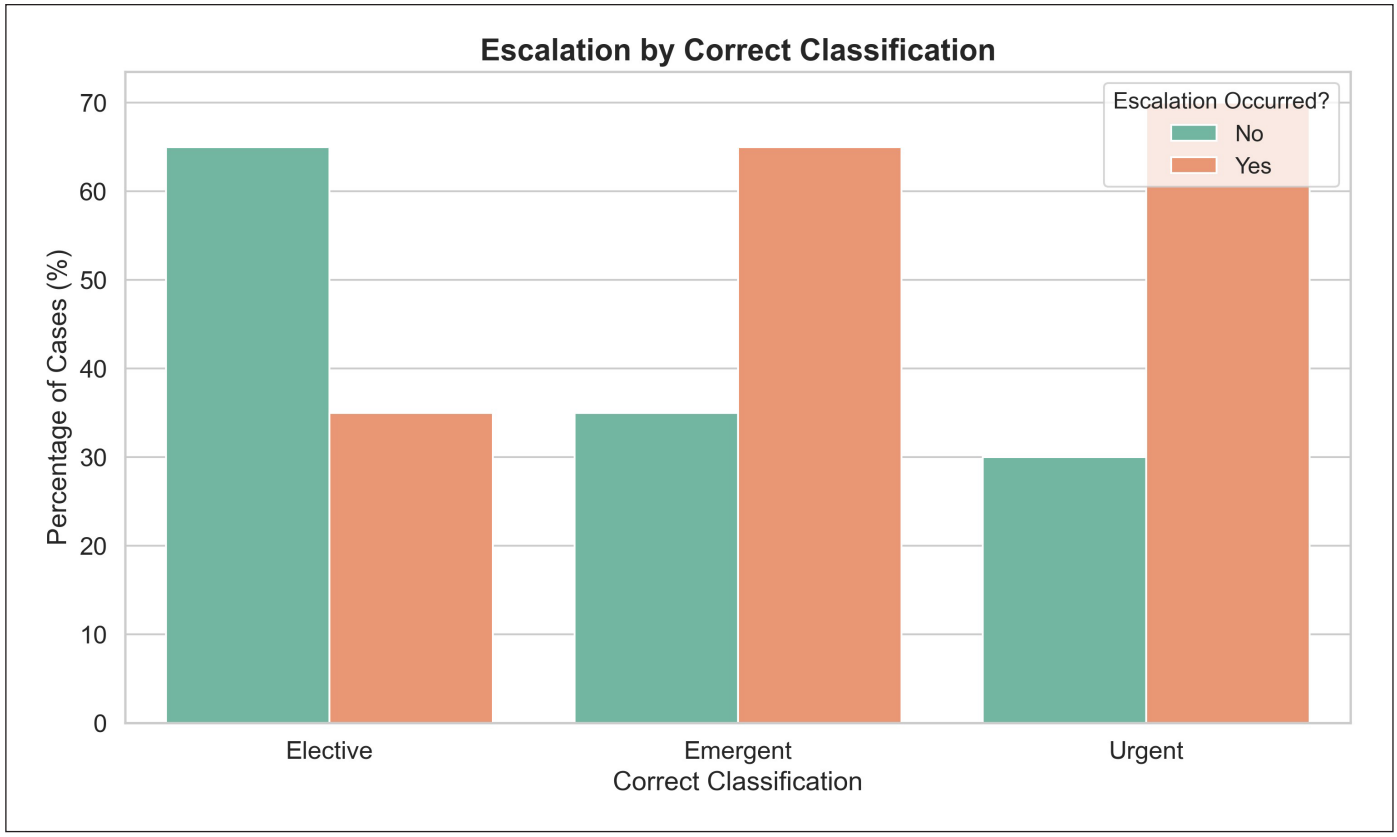

Figure 2. Escalation behavior divided based on correct or incorrect scenario classification. Among the scenarios correctly classified, escalation occurred in 65% and 70% of emergent and urgent cases, respectively. Only 35% of correctly classified elective cases were escalated. Overall, when classification was accurate, chatbots escalated more appropriately based on the given clinical urgency.

Table 3. Summary of Statistics by Classification

Classification	True negatives	False negatives	True positives	False positives	FNR (%)	PPV	NPV
Emergent	40	16	4	0	80.00	1	0.71
Urgent	23	3	17	17	15.00	0.50	0.88
Elective	36	2	18	4	10.00	0.82	0.95

Odds ratio for emergent cases is infinite due to the absence of false positives in that category, not because all emergent cases were correctly identified. This reflects a zero denominator in the odds ratio formula, which occurs when there are no incorrect emergent classifications. FNR, false negative rate; NPV, negative predictive value; PPV, positive predictive value.

chatbots was 2.7. In 48.3% of scenarios, chatbots terminated the conversation before escalating, and in the remaining 51.7% of interactions, the chatbots either (1) referred the patient to a human provider or (2) sent relevant information directly to one.

Disclaimers and Warnings

In an effort to give credit where it may be due, the plastic surgery websites were checked to see if any included disclaimers noting that chatbots ought not to be employed for emergent or urgent concerns, and that patients should consult a real provider as needed. Upon investigation, only 1 of the 20 websites included a disclaimer stating that the messaging platform "cannot replace in-person medical care." Five other clarifying statements were found, but these were standard legal or privacy notices indicating that, by employing the chatbot, the user agreed to the site's privacy policy. These statements did not clarify the chatbots' limited clinical capacities or discourage over-reliance on them in medical emergencies.

Subgroup Analysis

Subgroup analysis was conducted across 2 dimensions: chatbot type and marketed function. When stratified by type, rule-based (button/flow) chatbots had the lowest error rate (41%), whereas hybrid models and artificial intelligence [AI]/natural language processing [NLP]–based systems demonstrated higher error rates, exceeding 65% (Figure 4). In contrast, subgrouping by the marketed function of the chatbot indicated that chatbots designed for live chat triage achieved higher CUQ (usability) scores on average than those currently marketed for FAQ/education and appointment scheduling purposes (Figure 3A). Structural design and intended function of the chatbot, therefore, are influential factors in the determination of chatbot

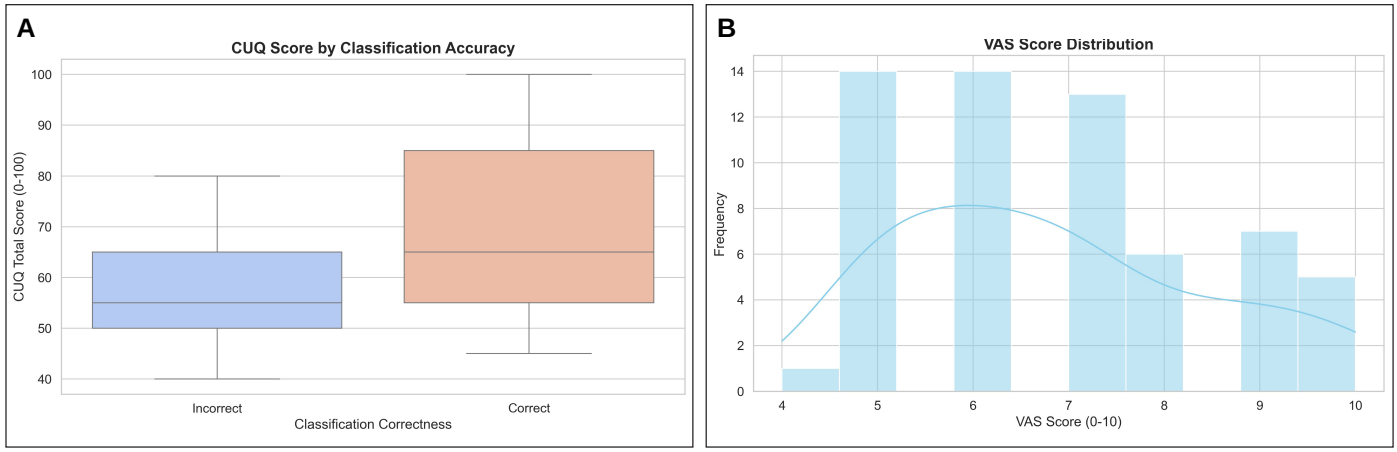

Figure 3. User experience scores based on chatbot performance. (A) CUQ (chatbot usability questionnaire) scores stratified by classification correctness. (B) VAS (visual analog scale) scores stratified by classification correctness.

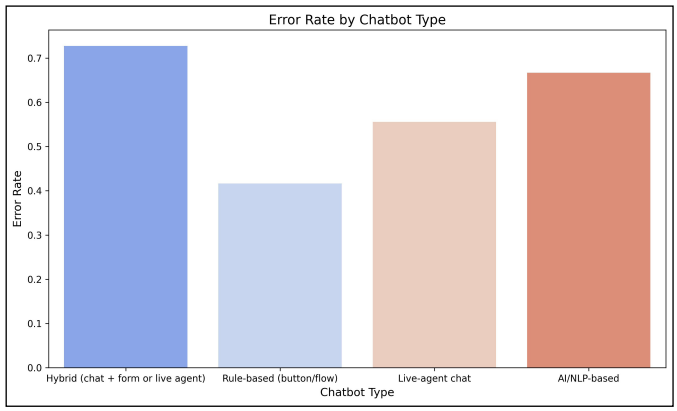

Figure 4. Error rates stratified by chatbot "type." Rule-based chats had the lowest error rates (42%), whereas hybrid models (72%) and AI/NLP-based systems (68%) had the greatest error rates. AI, artificial intelligence; NLP, natural language processing.

Table 4. Chatbot Responses

Theme	Responses (%)	Explanation
Clinical escalation	11 (30.55)	Responses urge the user to contact a medical provider or take urgent action.
Educational/ informational	11 (30.55)	Replies explain procedures, expected outcomes, or recovery guidance.
Templated advice	7 (19.44)	Frequent "stock phrases" commonly found in medical chatbot responses.
Administrative/ scheduling	7 (19.44)	Logistics-focused messages, including appointment bookings and contact information requests.

The majority of responses were categorized as clinical escalation, followed closely by educational and informational.

performance and usability. There was also a clear connection between user satisfaction and technical performance of the chatbot itself, as shown by VAS (visual analog scale) score distributions, in which chatbots with lower error rates and more appropriate triage behaviors were correlated with a higher user satisfaction score (Figure 3B).

Thematic Analysis

Among the chatbot replies, 36 unique chatbot responses were identified and clustered. Unsupervised K-means clustering, which was applied to TF-IDF–vectorized text, identified 4 dominant themes, which were then clustered: (1) clinical escalation (11 responses), (2) educational/informational, (3) templated posttreatment advice, and (4) administrative/scheduling. Table 4 shows the proportion of the responses and provides a brief description of the themes themselves.

The above themes were then visualized in a principal component analysis with TF-IDF matrices, with the first 2 principal components plotted to visualize linguistic structure across themes. Templated advice was highly cohesive, clustering in the upper right quadrant, and clinical escalation was clustered in the far left, exhibiting consistent language and lexical uniformity in these categories. The educational and administrative replies were distinct, but had greater dispersion on the whole. The educational and administrative categories also overlapped, indicating shared chatbot vocabulary between the 2 themes (Figure 5).

The word cloud shows that terms such as "procedure," "consultation," "augmentation," "phone," and "assistant" were among the most frequent in the chatbot logs (Figure 6). These words, alongside the most popular theme categories, as determined by principal component analysis (PCA), indicated that a majority of chatbot responses were focused on initiating administrative workflows while delivering templated procedural content.

DISCUSSION

The use of artificial intelligence chatbots on patient-forward websites is an evolving frontier in the plastic surgery space. Machine learning [ML]–integrated user interfaces offer on-demand patient interaction and clinical advice, while also streamlining surgical team support workflows.[16] The findings of this study demonstrate this notable strength of chatbots, namely, their capability of providing elective inquiries and reducing logistical burdens. Nevertheless, the performance of AI chatbots in urgent and emergent contexts was suboptimal and inconsistent. These tools may yield up-front cost savings for healthcare systems, but pose genuine risks to overall patient care, particularly in the case that escalation to a human provider is delayed, or fails altogether. These results have broad implications for

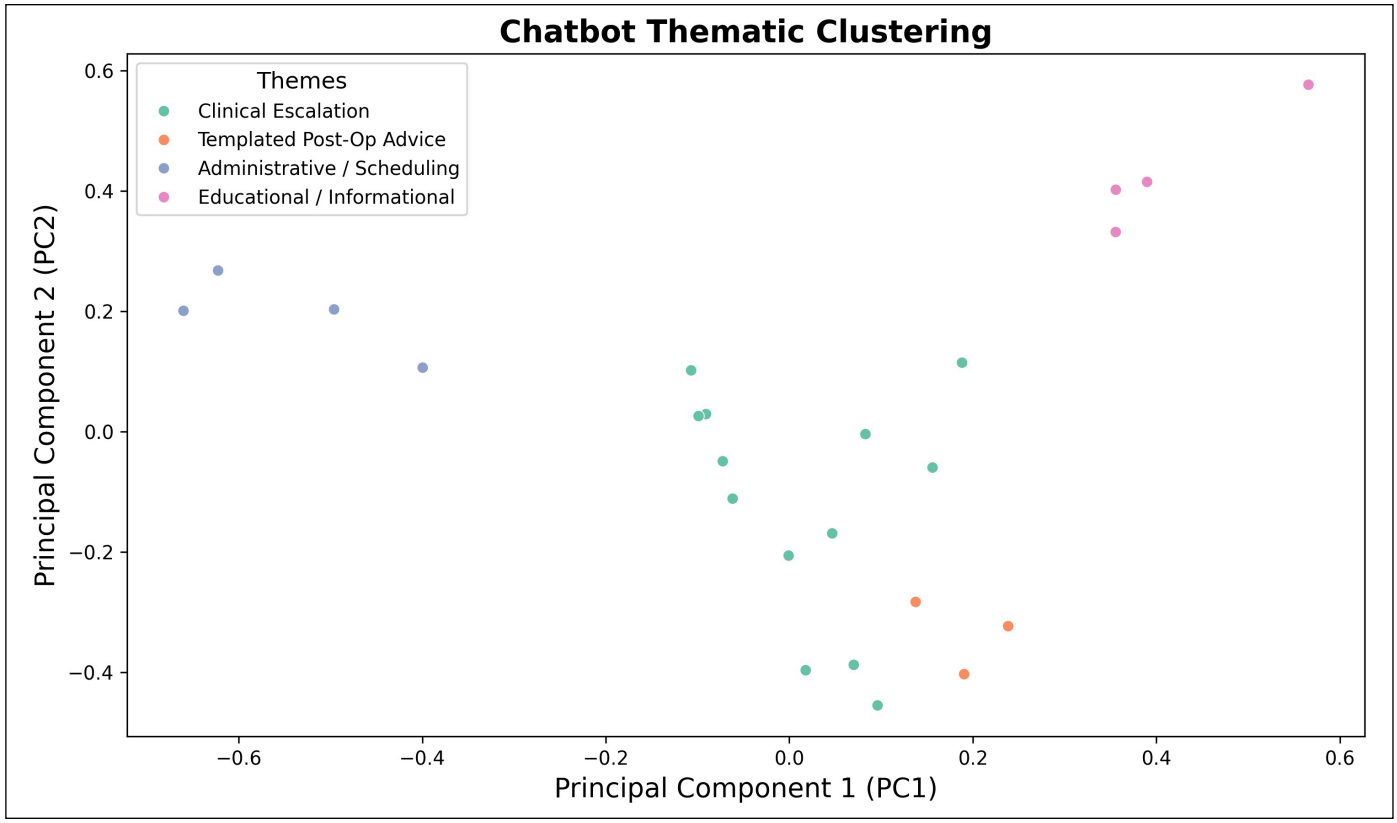

Figure 5. Two-dimensional scatter plot (via truncated SVD) demonstrating thematic clustering of chatbot replies. Each point represents a chatbot response, with clustering groups indicative of distinct linguistic patterns. Thematic clusters were assigned by K-means analysis, with green representing administrative/scheduling, orange clinical escalation, blue templated operative advice, and pink educational/informational. SVD, singular value decomposition.

plastic surgery and other specialties applying AI to clinical triage. Chatbot triage classification accuracy was highest for elective cases (90%), with a negative predictive value of 0.95, indicating that one can be most confident in chatbots for nonurgent matters.

This finding, supported by qualitative analysis of chatbot transcripts, demonstrates that AI chatbots are well-suited for administrative and consultative purposes, including scheduling procedures, providing educational material, and filtering nonurgent traffic. Unfortunately, this same confidence cannot be conferred to chatbot classification of emergent and urgent scenarios. In spite of the fact that a majority of chatbots are advertised to serve elective customer service purposes by their engineers, many practices implement these tools beyond their prescribed scope. Evidenced by drop-offs in correct classification rates for urgent and emergent cases, 20.0% and 34.0% respectively, urgent cases are commonly not escalated promptly or are handled with generic agentic language. The sensitivity data reinforce this critical concern about patient safety in the context of clinical triage, particularly when chatbots are employed to assess emergent conditions beyond their intended administrative or elective scope. That is, 4 out of 5 emergent cases were missed entirely, as compared to urgent and elective cases with sensitivities of 85% and 90%, respectively. This statistic highlights the substantial limitation of the chatbot models on plastic surgery websites to accurately identify scenarios in which patients may be in critical danger.

Although 72.2% of misclassified interactions were eventually escalated, delayed triage in high-risk cases, such as postoperative complications, indicates a potential for direct patient harm. The findings of our study consequently raise significant ethical and safety concerns, particularly in the realm of elective plastic surgery. Postoperative complications that escalate quickly require reliable escalation pathways.

The substantial accuracy in handling elective inquiries reflects the current absolute clinical and economic utility of chatbots in the elective surgery space. In this study, these chatbot tools were particularly effective in processing elective queries with minimal error and high satisfaction scores. Elective consultations, which often form the bulk of patient traffic in aesthetic surgery, benefit from asynchronous, automated interaction as highlighted in many previous discussions.[17] By managing this influx efficiently, chatbots hold the potential to reduce staff workload, improve patient onboarding, and offer a full-time digital front door to the nation's elite practices. In a market-driven specialty like plastic surgery, in which lead nurturing and patient experience are paramount, this feature alone justifies chatbot implementation, that is, when deployed within their realm of utility.

The qualitative review and thematic analysis performed highlights frequent reliance on templated, administrative responses in complex scenarios, which contributed to poor recognition of emergent issues, as well as poor overall patient satisfaction. This signals a design limitation in chatbot training data sets, in which language models may lack exposure to clinical emergencies or rare complications, especially in cases of elective surgery that deviate from the typically encountered patterns. To mitigate this, chatbot designers must integrate multidimensional training sets that incorporate not only medical knowledge but also aesthetic judgment heuristics and postoperative monitoring protocols. Future directions may include development of specialty-trained small language models (SLMs), which are compact AI tools tailored to specific clinical domains, including niche, aesthetic surgical concerns. However, for now, chatbots should default to escalation for any complications, because these tools currently lack the required data and training to accurately triage emergencies.

Figure 6. Word cloud based on logged text files of interactions with chatbots; the size of terms and phrases are correlated with the frequency with which they appeared in the logs. Procedural language (eg, "process," "augmentation," "breast augmentation") was repeated most often, and administrative (eg, "appointment," "phone") and templated postoperative advice (eg, "contact surgeon," "assistance") were also relatively popular.

Our study has several limitations, many due to the inherent design of the observational cross-sectional analysis. Namely, there is potential selection bias in identifying the surgery websites for inclusion, despite reliance on non–location-based SEO rankings. Furthermore, the study was limited to publicly accessible chatbot tools, which may not be representative of the full landscape of chatbot–patient interaction. The standardized clinical scenarios, as well, were helpful in interpreting chatbot performance but failed to capture the complexity of a true patient inquiry. It could also be argued that the sample size of the study was small at $n = 60$, which limited the power of our subgroup analyses. Furthermore, a more extensive data set would have produced more robust associations between chatbot function, design, and performance. Chatbot responses may have also evolved since the time of this study, because AI is constantly improving. We believe that a follow-up study with real patient inquiries and longitudinal evaluation of changes in CUQ and VAS metrics across plastic surgery platforms would provide valuable insight into the continuing evolution of chatbot practices on medical websites.

CONCLUSIONS

In the aesthetic surgery space, AI chatbots are highly effective at addressing elective inquiries and streamlining consultation logistics. The performance of these same chatbots for urgent and emergency clinical scenarios is suboptimal, with unignorable limitations in their accuracy and escalation capabilities. The results of this study therefore support that the present evolution of AI chatbot technology serves administrative and front-end patient engagement, and simultaneously underscore that they are ill-equipped for clinical advice. If clinicians want to deploy chatbots for the latter purpose, the technology must be developed to incorporate escalation safeguards and specialty-specific trained models like small language models (SLMs). Until these models are developed and deployed, chatbots should be limited to administrative, elective-oriented tasks, serving as an adjunct and not a replacement for physician oversight.

Supplemental Material

This article contains supplemental material located online at https://doi.org/10.1093/asj/sjaf123.

Disclosures

The authors declared no potential conflicts of interest with respect to the research, authorship, or publication of this article. During the preparation of this manuscript, the authors utilized ChatGPT (developed by OpenAI) to identify any possible grammar and spelling inconsistencies. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final version of the work.

Funding

The authors received no financial support for the research, authorship, or publication of this article.

REFERENCES

1. Uzoka NA, Cadet NE, Ojukwu NPU. Leveraging AI-powered chatbots to enhance customer service efficiency and future opportunities in automated support. *Comput Sci & IT Res J*. 2024;5:2485-2510. doi: 10.51594/csitrj.v5i10.1676
2. Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med*. 2024;30:2878-2885. doi: 10.1038/s41591-024-03148-7
3. Bernstein IA, Zhang Y, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6:e2330320. doi: 10.1001/jamanetworkopen.2023.30320
4. Turea M. *Chatbots on Medical Websites: why Your Patients Like Them*. Digital Authority Partners; 2024. Accessed April 21, 2025. https://www.digitalauthority.me/resources/chatbots-medical-websites-patients-like-them/

5. Jones C. *Chatbots for Service Desks: Boosting ITSM Efficiency and Satisfaction*. TOPdesk—EN; 2025. Accessed April 21, 2025. https://www.topdesk.com/en/blog/chatbot-service-desk/
6. Mierzwa SJ, Souidi S, Conroy T, Abusyed M, Watarai H, Allen T. On the potential, feasibility, and effectiveness of chat bots in public health research going forward. *Online J Public Health Inform*. 2019;11:e4. doi: 10.5210/ojphi.v11i2.9998
7. George-Todorov. *70 Fantastic Chatbot Statistics, Trends and Facts for 2024*. Create & Grow; 2024. Accessed April 21, 2025. https://createandgrow.com/chatbot-statistics/
8. Fokina M. *The Future of Chatbots: 80+ Chatbot Statistics for 2025*. Tidio; 2024. Accessed April 21, 2025. https://www.tidio.com/blog/chatbot-statistics/
9. 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;183:589. doi: 10.1001/jamainternmed.2023.1838
10. Brandtzaeg PB, Følstad A. *Why People use Chatbots*. In: Kompatsiaris I, Cave J, Satsiou A, Carle G, Passani A, Kontopoulos E, Diplaris S, McMillan D, eds. Internet Science. INSCI 2017. *Lecture Notes in Computer Science*, vol. 10673. Springer. doi: 10.1007/978-3-319-70284-1_30
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2:e0000198. doi: 10.1371/journal.pdig.0000198
12. DSCCP Consultant. *AI Productivity Tools for Plastic Surgery Practices*. Specialist Practice Excellence; 2025. Accessed April 21, 2025. https://specialistpracticeexcellence.com.au/blog/ai-productivity-tools-for-plastic-surgery-practices/
13. AD Vital. *What Role Does AI Play in Securing Cosmetic Surgery Leads?* AD Vital; 2025. Accessed April 21, 2025. https://advitalmd.com/ai-cosmetic-surgery-leads-2/#:~:text=With%20AI,before%20stepping%20into%20a%20clinic
14. Rao A, Pang M, Kim J, et al. Assessing the utility of CHATGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. 2023;25:e48659. doi: 10.2196/48659
15. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential diagnosis lists generated by generative pre-trained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. 2023;20:3378. doi: 10.3390/ijerph20043378
16. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ*. 2023;23:689. doi: 10.1186/s12909-023-04698-z
17. American Society of Plastic Surgeons. *American Society of Plastic Surgeons Reveals 2022s Most Sought-After Procedures*. American Society of Plastic Surgeons; 2023. Accessed April 21, 2025. https://www.plasticsurgery.org/news/press-releases/american-society-of-plastic-surgeons-reveals-2022s-most-sought-after-procedures