LESS

Original Article

Laparosc Endosc Surg Sci 2026;33(2):61-66
DOI: 10.14744/less.2026.50374

Limitations and practical role of large language models in preoperative patient education: An expert-based assessment

Ilkay Halıcıoğlu, Ozcem Ofkeli

Department Of General Surgery, Diyarbakır Gazi Yaşargil Training and Research Hospital, Diyarbakır, Türkiye

ABSTRACT

Introduction: To evaluate the quality and educational value of information generated by large language models (LLMs) for use in preoperative counselling with patients undergoing sleeve gastrectomy.

Materials and Methods: Two LLMs were used to generate responses to common patient questions. Bariatric surgeons then evaluated the responses using a Likert scale to assess their accuracy, clarity and usefulness.

Results: Both models mostly produced accurate and clear information on procedure details, though individual risk communication was sometimes limited.

Conclusions: LLMs can provide supplementary information for preoperative education, but should not replace discussions with physicians. During preoperative counselling, clinicians should guide patients in interpreting AI-generated content.

Keywords: Artificial intelligence, patient education, large language models, bariatric surgery, healthcare communication

Introduction

Obesity is a significant public health problem with a rapidly increasing prevalence worldwide. It is a leading cause of both mortality and morbidity. According to the World Health Organization, there were more than 650 million obese individuals worldwide as of 2022.[1] The first-line approaches to treating obesity are lifestyle changes and medical treatments. However, when these interventions fail to yield satisfactory results, bariatric surgery becomes a viable option. Among these surgical methods, sleeve gastrectomy is one of the most commonly performed procedures.[2]

The preoperative information process is important for motivating individuals considering bariatric surgery, ensuring postoperative compliance, and achieving successful outcomes.[3] Traditionally, this process has been carried out through face-to-face information sessions, brochures, or educational videos. However, the growing workload of healthcare systems and the increasing need for personalized counseling have sparked interest in alternative information methods.

In this context, the use of artificial intelligence (AI)-based systems in patient education and healthcare information has increased.[4] Large language models (LLMs) and chatbots have the potential to improve health literacy by providing patients with quick, accessible information.[5] However, the content, accuracy, and adequacy of AI-supported information regarding patient safety have yet to be

Received: 08.02.2026 **Revision:** 06.04.2026 **Accepted:** 20.04.2026
Correspondence: Ilkay Halıcıoğlu, Department Of General Surgery, Diyarbakır Gazi Yaşargil Training and Research Hospital, Diyarbakır, Türkiye
e-mail: ilkayhlcgl@gmail.com

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

clearly established. In surgeries such as gastric bypass, which require multidisciplinary preparation and information, it is debatable whether AI applications can replace traditional clinical information processes.[6]

Although several previous studies have examined the use of large language models (LLMs) in educating patients about bariatric surgery, most have focused on a single platform or different versions of the same model. This study provides a blinded, head-to-head comparison of two widely used consumer-facing LLM ecosystems (ChatGPT and Gemini), using an identical set of questions, a standardised prompt structure and an expert-based scoring framework. Beyond overall performance scores, the study identifies specific areas where both models demonstrate consistent limitations, particularly in personalized risk assessment and individualized nutritional counselling. These findings offer incremental value by defining the practical boundaries for safely integrating LLMs as supplementary tools in the preoperative education of bariatric patients, rather than as standalone sources of information.

This study aimed to evaluate and compare the performance of two large language models when answering preoperative patient questions about sleeve gastrectomy. The focus was on accuracy, comprehensiveness and clarity, as assessed by experts.

Materials and Methods

This study was designed to evaluate patient-facing information generated by two large language models in response to common preoperative questions about sleeve gastrectomy.

A structured questionnaire consisting of 21 questions was developed to reflect the type of information commonly requested by patients prior to undergoing laparoscopic sleeve gastrectomy. The questions were generated based on a combination of clinical experience, frequently asked patient inquiries during preoperative bariatric surgery consultations, and a focused review of the literature on patient education in bariatric surgery. All questions were posed in Turkish.

These questions were then input into two widely used artificial intelligence chatbots—ChatGPT (GPT-4.0 version) and Google Gemini (as available in April 2024). The responses generated by each chatbot were recorded verbatim for subsequent evaluation. To evaluate the quality of these responses, eight board-certified surgeons with experience in bariatric surgery—including general surgeons, gastrointestinal surgeons, and surgical oncologists—were recruited as expert raters. For the purposes of this study, an 'expert bariatric surgeon' was defined as a general surgeon who is certified by a board and is actively practising bariatric surgery within a multidisciplinary bariatric programme. All participating experts had a minimum of five years' experience in bariatric surgery and had performed more than 200 bariatric procedures cumulatively. In addition to their surgical work, all evaluators routinely participated in preoperative patient counselling and postoperative follow-up, ensuring they were familiar with patient education needs and commonly asked clinical questions. The experts were blinded to the identity of the AI models providing each set of answers. They were instructed to evaluate each individual answer according to three key criteria:

1. **Accuracy** (clinical correctness of the information),
2. **Comprehensiveness** (completeness and relevance of the content),
3. **Clarity** (ease of understanding and appropriateness for a layperson).

Each criterion was rated on a **5-point Likert scale**, where:

- **1 = Very Poor**
- **2 = Poor**
- **3 = Fair**
- **4 = Good**
- **5 = Excellent**

The total score for each response was calculated as the sum of the three ratings (range: 3 to 15 points per response). The aggregated scores across all 21 questions and 8 raters were then compared between the two AI models using **paired t-tests** to evaluate statistical significance.

The study was approved by the local institutional ethics committee (7 February 2025 Approval number: 332) and conducted according to Helsinki Declaration

A total of 21 standardized patient questions related to sleeve gastrectomy were posed to both AI models. The full list of these questions is provided in the Appendix.

Results

A total of 21 questions relevant to sleeve gastrectomy were evaluated using responses generated by two large language models: ChatGPT (GPT-4.0) and Gemini (as of April 2024). Each response was assessed independently by eight expert surgeons using a 5-point Likert scale across three dimensions: Accuracy, comprehensiveness, and clarity. For each answer, a total score ranging from 3 to 15 was calculated.

Overall Performance

The mean total score across all questions was 14.21 (±0.31) for ChatGPT and 13.79 (±0.39) for Gemini. A paired t-test revealed a statistically significant difference between the two models (t=2.098, p=0.049), suggesting that there is a slight difference in the perceived quality of responses from ChatGPT and Gemini among experts. The mean total scores for ChatGPT and Gemini across 21 preoperative patient questions are summarised in Table 1.

When asking questions to AI robots, no guidelines or textbooks were used as sources, and questions were asked directly and clearly. The aim here was for the AI to provide answers appropriate to the average level of awareness and knowledge of patients and their relatives who have direct access to the internet. Experts also took this into consideration when evaluating the answers. Based on feedback from many experts who participated in the study, we found that Chat GPT 4.0 uses a more informal language, while the Google Gemini application uses a more formal language.

To ensure methodological consistency and reproducibility, both large language models were queried using the same prompt structure and set of 21 questions in the same order. No follow-up prompts, clarifications or iterative refinements were applied. All responses were recorded verbatim at a single time point in February 2025, without any post-processing or manual modification. This approach aimed to simulate how an average patient or relative might interact directly with publicly accessible AI tools. The full prompt template used in this study can be found in the Supplementary Materials.

Question-by-Question Analysis

A detailed analysis for each individual question showed that ChatGPT scored higher than Gemini in 14 out of 21 questions, while Gemini outperformed ChatGPT in 7 questions. For each question, a separate paired t-test was conducted using expert ratings. The following questions demonstrated statistically significant differences ($p<0.05$) between the models:

Q6: What preoperative investigations are required before sleeve gastrectomy?

Q7: How long is the hospital stay after surgery?

Q16: How frequently are postoperative follow-up visits required?

Q18: Will I experience excess skin after weight loss, and how can it be managed?

- **Q6** – Gemini significantly outperformed ChatGPT (p=0.0138)
- **Q7** – ChatGPT significantly outperformed Gemini (p=0.0331)
- **Q16** – ChatGPT significantly outperformed Gemini (p=0.0203)
- **Q18** – ChatGPT significantly outperformed Gemini (p=0.0021)

For the remaining questions, although score differences existed, they did not reach statistical significance. Comparative results for each question are shown in Table 1.

Trend Observation

Despite minor question-level variations, ChatGPT consistently received high scores in clarity and accuracy, while Gemini showed relatively more variability across raters. No critical content errors were reported in either model, though the breadth and phrasing of responses varied noticeably.

Although the difference in overall scores between ChatGPT and Gemini was statistically significant (14.21 vs. 13.79 on a 15-point scale; p=0.049), the magnitude of this difference was small. Given the high overall scores observed for both models, it is likely that a ceiling effect is limiting the practical interpretability of minor numerical differences. Therefore, the results should be interpreted as indicating relative performance trends rather than superiority that is decisive in clinical or educational contexts.

Discussion

This study evaluated and compared the quality of responses generated by two large language models (LLMs)

Table 1. Model comparison by question with p-values

Question	ChatGPT Avg	Gemini Avg	Difference	Higher Model	p	Significant (p<0.05)
Q1	14.250	14.125	0.125	ChatGPT	0.8153	No
Q2	14.375	13.625	0.750	ChatGPT	0.3645	No
Q3	13.625	14.625	-1.000	Gemini	0.0676	No
Q4	14.500	13.875	0.625	ChatGPT	0.1395	No
Q5	12.750	13.375	-0.625	Gemini	0.4051	No
Q6	12.875	14.500	-1.625	Gemini	0.0138	Yes
Q7	14.500	13.000	1.500	ChatGPT	0.0331	Yes
Q8	12.875	13.000	-0.125	Gemini	0.8952	No
Q9	14.750	14.500	0.250	ChatGPT	0.5983	No
Q10	14.625	13.000	1.625	ChatGPT	0.4331	No
Q11	14.125	13.750	0.375	ChatGPT	0.5040	No
Q12	15.000	14.625	0.375	ChatGPT	0.3506	No
Q13	14.125	14.375	-0.250	Gemini	0.4512	No
Q14	14.125	12.500	1.625	ChatGPT	0.1288	No
Q15	14.625	13.500	1.125	ChatGPT	0.0514	No
Q16	15.000	13.625	1.375	ChatGPT	0.0203	Yes
Q17	14.250	14.000	0.250	ChatGPT	0.6263	No
Q18	15.000	13.000	2.000	ChatGPT	0.0021	Yes
Q19	15.000	14.250	0.750	ChatGPT	0.1705	No
Q20	13.625	13.875	-0.250	Gemini	0.7406	No
Q21	14.500	14.500	0.000	Equal	1.0000	No

This table presents the average total scores for each question evaluated by experts for both ChatGPT and Gemini. It also includes the score difference, statistical significance (p-value) from paired t-tests, and whether the difference is significant (p<0.05)

— ChatGPT (GPT-4.0) and Gemini — to common preoperative questions asked by patients about sleeve gastrectomy. Both models demonstrated a high overall level of accuracy, clarity and comprehensiveness using a structured expert-based scoring system, supporting their potential role as supplementary sources of patient information.[2,7,8]

Although ChatGPT achieved slightly higher overall scores than Gemini, this difference was small and unlikely to be clinically significant. These findings suggest that both models can deliver reliable, understandable information for standard, non-individualised patient education. Similar trends have been reported in previous studies evaluating AI-generated responses in bariatric surgery, where GPT-based models have performed well in informational and counselling contexts.[7,8]

However, a question-level analysis revealed that performance varied depending on the type of information requested. Both models performed well when addressing general procedural explanations and postoperative expectations, but showed consistent limitations when responding to questions requiring personalised risk assessment or nutritional counselling. Questions such as 'Is this surgery risky for me?' and 'Will I need additional nutritional support?' were rated relatively poorly, regardless of the model used. This reflects an inherent limitation of current LLMs, which lack access to patient-specific clinical data and the capacity for individualised medical judgement.[3,6]

Another important observation was the difference in communication style between the models. Expert evaluators noted that ChatGPT tended to use a more informal, patient-friendly tone, whereas Gemini responses were generally more formal. While both styles may be appropriate in different contexts, clarity and accessibility are particularly important in patient education. This may partially

explain why ChatGPT received higher clarity scores for certain questions.[7,9]

Several limitations of this study should be acknowledged. Firstly, the evaluation was based solely on expert opinion and did not include direct assessment of patient comprehension, health literacy or behavioural outcomes.[6] Secondly, although all evaluations were performed independently using a standardised framework, inter-rater reliability metrics were not formally calculated. Finally, the analysis was limited to responses generated at a single time point and in a single language, which may affect generalisability.

Recent studies have highlighted the expanding role of large language models in healthcare, particularly in patient education, clinical communication, and decision support. However, concerns remain regarding variability in accuracy, potential hallucinations, and the need for expert validation. Our findings are consistent with this growing body of literature, suggesting that while LLMs may provide useful supportive information, they should not replace clinician-guided counseling.[10-12]

Another important consideration is the language used during model interaction. Since all questions in this study were posed in Turkish, the findings may not be directly generalizable to responses generated in other languages. Differences in linguistic structure and medical terminology may influence model performance and should be considered in future studies.

Interestingly, statistically significant differences between the models were observed in specific domains, particularly preoperative evaluation (Q6), perioperative expectations (Q7), long-term follow-up (Q16), and aesthetic outcomes such as excess skin after weight loss (Q18). These findings suggest that model performance may vary depending on the nature of the question. While some questions require structured, guideline-based clinical knowledge, others involve patient-centered counseling and experiential explanations. This variability highlights that AI-generated responses may not be equally reliable across all aspects of bariatric patient education.

The question-specific variability observed in this study suggests that large language models may perform more consistently in general explanatory contexts than in areas requiring structured follow-up recommendations or nuanced clinical counseling. This finding has direct implications for the safe integration of LLMs into clinical practice, particularly in high-stakes decision-making contexts.

Conclusion

In this expert-based comparative assessment, both ChatGPT and Gemini demonstrated the ability to generate accurate, clear and comprehensive responses to common sleeve gastrectomy-related questions. Although ChatGPT showed slightly higher overall performance scores, the difference between the two models was modest and should not be interpreted as evidence of clear superiority in clinical or educational practice.

Importantly, both models consistently struggled with questions requiring individualised clinical judgement, risk stratification or personalised nutritional planning. These findings reinforce the idea that large language models should not be used as standalone tools for patient counselling or medical decision-making.

Practice implications

Large language models could support preoperative patient education by providing accessible, standardised information about bariatric surgery. However, they should be used in a complementary role only and must always be accompanied by direct physician–patient communication. Clinicians should actively guide patients in interpreting AI-generated information, clarify misconceptions and ensure that individualised risks and expectations are discussed during the consultation.

Appendix: *https://jag.journalagent.com/less/abs_files/LESS-50374/LESS-50374_(2)_LESS-50374_(0)_Appendix.pdf*

Disclosures

Ethics Committee Approval: The study was approved by the local institutional ethics committee (Date: 07/02/2025, No: 332)

Peer-review: Externally peer-reviewed.

Conflict of Interest: The authors declare that they have no financial conflicts of interest, sponsorships, consulting relationships, or institutional affiliations related to this study.

Funding: This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions: Concept – I.H.; Design – I.H.; Supervision – I.H.; Funding – I.H.; Materials – I.H.; Data Collection – I.H., O.O.; Analysis and/or interpretation – I.H.; Literature Search – I.H., O.O.; Writing – I.H.; Critical Review – I.H.

Acknowledgments: We would like to thank the general surgeons who contributed to the evaluation of the artificial intelligence systems included in this study. The participating surgeons volunteered to take part in all stages of the evaluation process and conducted their assessments independently. We also sincerely thank **Assoc. Prof. Kadri Çolakoğlu, MD**, for his valuable academic guidance and constructive feedback during the development of this study.

References

4. World Health Organization. Obesity and overweight. Geneva: World Health Organization; 2022.
5. Angrisani L, Santonicola A, Iovino P, Formisano G, Buchwald H, Scopinaro N. Bariatric surgery worldwide 2013. Obes Surg 2017;25(10):1822–32.
6. Mechanick JI, Youdim A, Jones DB, Garvey WT, Hurley DL, McMahon M, et al. Clinical practice guidelines for the perioperative nutritional, metabolic, and nonsurgical support of the bariatric surgery patient. Obesity 2013;21(Suppl 1):S1–27.
7. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.
8. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou C, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24–9.
9. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM, et al. Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2022;24(3):e30235.
10. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33(6):1790–6.
11. Samaan JS, Rajeev N, Ng WH, Srinivasan N, Busam JA, Yeo YH, et al. ChatGPT as a source of information for bariatric surgery patients: A comparative analysis of accuracy and comprehensiveness between GPT-4 and GPT-3.5. Obes Surg 2024;34(5):1987–9.
12. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? JMIR Med Educ 2023;9:e45312.
13. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on the United States Medical Licensing Examination. PLOS Digit Health 2023;2(2):e0000198.
14. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172–80.
15. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589–96.