Original Article

Cukurova Anestezi ve Cerrahi Bilimler Dergisi

Guideline Concordance of Large Language Models for ERAS Colorectal Surgery Recommendations: A Blinded, Clinician-Rated Comparison of Google Gemini and ChatGPT

Sezer Gökçen[1]

[1] Department of General Surgery, Ağrı Training and Research Hospital, Ağrı, Türkiye

***Correspondence:** Sezer Gökçen, MD; drsezergokcen@gmail.com

Abstract

Background: Large language models (LLMs) are increasingly used by clinicians and trainees for perioperative decision support, yet their alignment with Enhanced Recovery After Surgery (ERAS) recommendations remains uncertain. **Methods:** We converted the 2025 ERAS Society recommendations for elective colorectal surgery into a 52-question bank (preoperative n = 28, intraoperative n = 10, and postoperative n = 14). Each question was asked once, in a new chat, to Google Gemini and OpenAI ChatGPT (web interfaces; no follow-up prompts or regeneration; queries performed on 17 Feb 2026). Responses were blinded as A/B and independently scored by two clinicians for (i) guideline concordance on a 5-point Likert scale and (ii) safety risk (0 = none, 1 = potential harm, 2 = critical harm). Primary analysis used paired Wilcoxon signed-rank tests (Likert) and exact McNemar tests (any safety flag ≥1). Inter-rater agreement was estimated with quadratic weighted kappa (QWK). **Results:** A total of 208 ratings were generated (52 questions × 2 models × 2 raters). Mean Likert concordance was 4.49±0.62 for Gemini and 4.44±0.65 for ChatGPT (paired Wilcoxon p = 0.604). Any safety flag occurred in 20.2% (Gemini) and 15.4% (ChatGPT) of ratings (McNemar p = 0.383); no responses were rated as critical harm. Inter-rater agreement was lower for Gemini (QWK = 0.225) than for ChatGPT (QWK = 0.597). **Conclusions:** Both LLMs showed high overall concordance with ERAS colorectal recommendations, with no significant overall difference in scores. However, safety-relevant deviations were not rare, and rater agreement varied by model, highlighting the need for clinician oversight and standardized evaluation frameworks before bedside use.

Keywords: ERAS; colorectal surgery; large language model; ChatGPT; Gemini; patient safety

Received: 12.02.2026
Accepted: 09.06.2026
Published: 30.06.2026

1. Introduction

Enhanced Recovery After Surgery (ERAS) pathways are now central to modern colorectal perioperative care and are supported by continually updated, evidence-based recommendations from the ERAS Society[1,2]. At the same time, large language models (LLMs) have rapidly entered clinical and educational workflows, raising the possibility that surgeons and trainees will increasingly consult LLMs for guideline-oriented answers.

However, LLM outputs may be non-deterministic, can "hallucinate" unsupported content, and may omit key nuances, creating a potential patient-safety risk if used without oversight[3].

Recent reporting frameworks (e.g., TRIPOD-LLM and TRIPOD+AI) emphasize the need for transparent, reproducible evaluation of LLMs in healthcare, and studies have begun benchmarking LLM performance in surgical decision-making and multidisciplinary settings[4,5,6,7]. Nevertheless, data are limited regarding LLM concordance with ERAS colorectal recommendations, and direct comparisons across models using blinded, clinician-based rating systems remain scarce.

We performed a blinded, clinician-rated comparison of Google Gemini and OpenAI ChatGPT using a structured question bank derived from the 2025 ERAS Society colorectal surgery recommendations[1]. We assessed guideline concordance, safety risk, and inter-rater agreement.

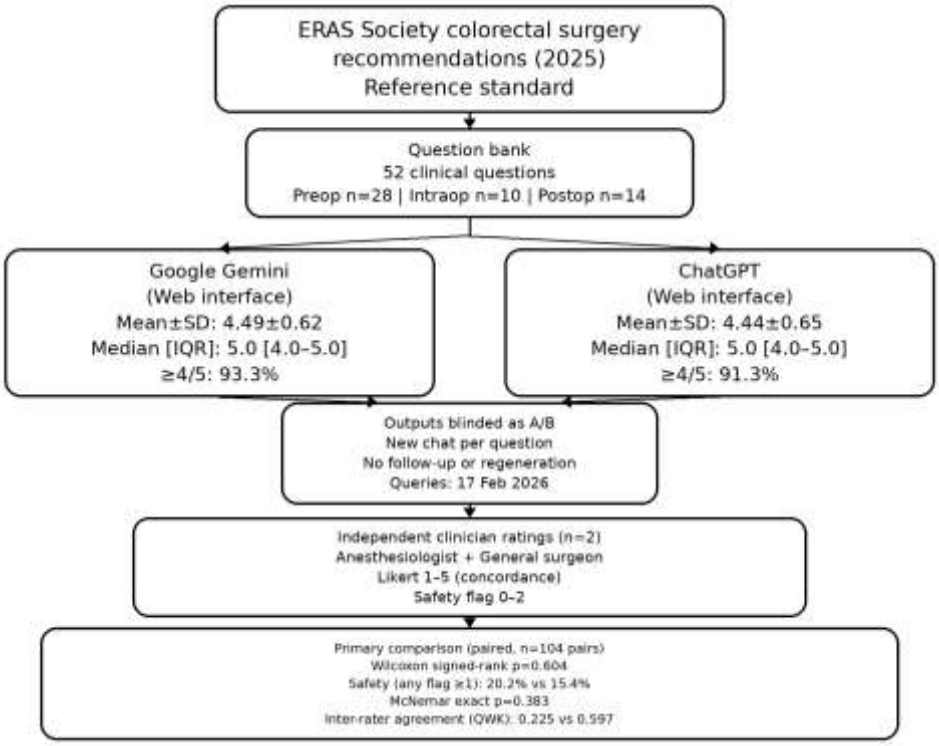

Figure 1. Study workflow and primary outcomes

2. Methods

Study design: cross-sectional evaluation of LLM outputs against a published guideline reference standard. No patient-level data, human participants, or identifiable information were used.

The 2025 ERAS Society recommendations for elective colorectal surgery served as the reference standard[1]. Each recommendation item was reviewed and independently converted into one or two clinical scenario questions designed to probe guideline-concordant decision-making. The resulting 52-question bank spanned the preoperative (n = 28), intraoperative (n = 10), and postoperative (n = 14) phases. The full question list is provided in Supplementary Material S1.

Each question was submitted in English using a standardized instruction prompt (Supplementary Material S2). Google Gemini and OpenAI ChatGPT were accessed via

their publicly available web interfaces. At the time of querying (17 February 2026), the model labels shown in the interfaces were recorded as Gemini 3 (Pro) and GPT-5.2, respectively. For each question, a new chat was started, and no follow-up prompts, regeneration, or manual editing of outputs were performed.

Model outputs were de-identified and labeled as A or B prior to evaluation. Two independent clinician raters (an anesthesiologist and a general surgeon) assessed each response. Guideline concordance was scored on a 5-point Likert scale (1 = incorrect or contradicts guideline, 5 = fully concordant and complete). Safety risk was scored as 0 (no safety concern), 1 (potential harm), or 2 (critical harm). The scoring rubric is included in Supplementary Material S3.

Primary comparisons used paired Wilcoxon signed-rank tests for Likert scores across matched question–rater pairs and exact McNemar tests for paired binary safety outcomes (any safety flag ≥1). Inter-rater agreement was summarized with quadratic weighted kappa (QWK) and Spearman correlation[8]. Analyses were performed in Python. Two-sided $p < 0.05$ was considered statistically significant; phase-specific analyses were exploratory.

3. Results

A total of 208 ratings were generated (52 questions × 2 models × 2 raters). Overall guideline concordance was high for both models (Table 1). Mean Likert scores were 4.49±0.62 for Gemini and 4.44±0.65 for ChatGPT with no statistically significant difference (paired Wilcoxon $p = 0.604$).

Any safety flag (≥1) occurred in 20.2% of Gemini ratings and 15.4% of ChatGPT ratings (exact McNemar $p = 0.383$). No responses were rated as critical harm (safety flag = 2). Inter-rater agreement differed by model: QWK was 0.225 for Gemini versus 0.597 for ChatGPT.

In phase-stratified analyses (Table 2), Gemini scored slightly higher in the preoperative domain, whereas ChatGPT tended to score higher in the postoperative domain. Topics with the largest between-model differences (≥0.5 absolute) are shown in Table 3. Per-question differences are shown in Figure 2.

Table 1. Overall guideline concordance and safety ratings (pooled across two clinician raters)

Model	n	Mean±SD	Median [IQR]	Likert ≥4/5	Safety flag ≥1	Critical harm
Google Gemini	104	4.49±0.62	5.0 [4.0–5.0]	97 (93.3%)	21 (20.2%)	0 (0.0%)
ChatGPT	104	4.44±0.65	5.0 [4.0–5.0]	95 (91.3%)	16 (15.4%)	0 (0.0%)

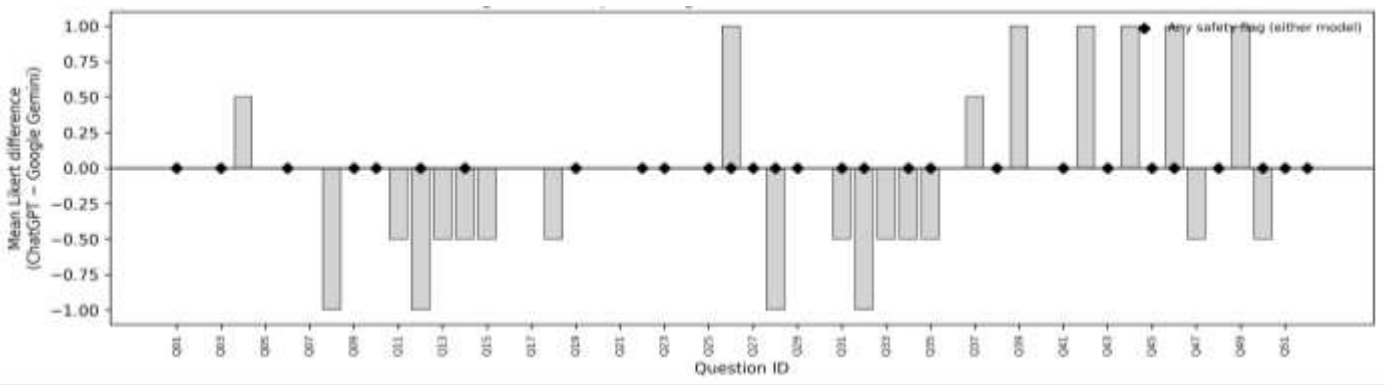

Figure 2. Per-question mean concordance differences (ChatGPT − Google Gemini). Diamonds mark questions for which at least one rater assigned any safety flag (≥1) to either model.

Table 2. Phase-stratified performance and paired comparisons (exploratory; p-values are paired tests within each phase)

Phase	Gemini Mean±SD	Gemini Median [IQR]	Gemini ≥4/5	Gemini safety ≥1	ChatGPT Mean±SD	ChatGPT Median [IQR]	ChatGPT ≥4/5	ChatGPT safety ≥1	p (Likert)	p (safety)
Preoperative	4.57±0.53	5.0 [4.0–5.0]	55/56 (98.2%)	8/56 (14.3%)	4.43±0.68	5.0 [4.0–5.0]	50/56 (89.3%)	8/56 (14.3%)	0.046	1.000
Intraoperative	4.65±0.49	5.0 [4.0–5.0]	20/20 (100.0%)	5/20 (25.0%)	4.40±0.75	5.0 [4.0–5.0]	17/20 (85.0%)	4/20 (20.0%)	0.059	1.000
Postoperative	4.21±0.79	4.0 [4.0–5.0]	22/28 (78.6%)	8/28 (28.6%)	4.50±0.51	4.5 [4.0–5.0]	28/28 (100.0%)	4/28 (14.3%)	0.068	0.219

Table 3. Topics with the largest between-model differences in mean concordance scores (absolute difference ≥0.5, top 12)

QID	Domain	ERAS Topic	Gemini mean	ChatGPT mean	Difference
Q44	Postop	Postoperative fluids	4.0	5.0	+1.0
Q46	Postop	Urinary drainage	4.0	5.0	+1.0
Q28	Preop	Preoperative fluids	4.5	3.5	−1.0
Q26	Preop	Antibiotics and skin preparation	3.5	4.5	+1.0
Q39	Postop	Nasogastric tube (NGT)	4.0	5.0	+1.0
Q42	Postop	Normoglycemia	3.5	4.5	+1.0
Q12	Preop	Anemia treatment	4.5	3.5	−1.0
Q32	Intraop	Normothermia	4.5	3.5	−1.0
Q08	Preop	Smoking cessation	4.5	3.5	−1.0
Q49	Postop	Postoperative analgesia	4.0	5.0	+1.0
Q47	Postop	Prevention of ileus	5.0	4.5	−0.5
Q04	Preop	Preoperative optimization	4.5	5.0	+0.5

4. Discussion

In this blinded evaluation against the 2025 ERAS colorectal recommendations, both Google Gemini and ChatGPT achieved high overall guideline concordance, and we observed no significant difference in overall Likert scores. These findings align with reports that LLMs can provide guideline-adjacent responses in structured surgical tasks[6,7,9,10].

Safety-relevant deviations were not rare (15–20% of ratings), even though none were considered critical harm. Potential-harm deviations—such as omission of key contraindications, oversimplification of fluid or analgesia recommendations, or premature generalization to high-risk populations—could affect outcomes if acted upon without clinical judgment. This supports calls for mitigation strategies and human oversight for clinical decision support use[3,4].

Inter-rater agreement was substantially higher for ChatGPT than for Gemini in this dataset. This may reflect differences in response structure or clarity rather than purely factual accuracy. ChatGPT's more consistently structured responses—typically organized into labelled sections—may have facilitated more uniform scoring, whereas Gemini's more varied output format may have introduced greater interpretive divergence between raters.

An important methodological consideration is output variability inherent to LLMs. Because each question was submitted only once per model, the observed responses represent a single sample from a stochastic output distribution. Repeated querying of the same question can yield meaningfully different responses in terms of content, emphasis, and completeness. Consequently, the concordance and safety scores reported here should be interpreted as a snapshot rather than a stable characterization of each model's performance. Future studies would benefit from repeated querying protocols or temperature-controlled API access to quantify within-model variability alongside between-model differences.

High aggregate concordance does not preclude clinically meaningful errors at the individual question level: a model that scores 4.4–4.5 on average may still produce guideline-discordant or potentially harmful responses for specific topics—as illustrated by the 15–20% safety flag rate. This pattern suggests that LLMs may be well-suited as supplementary educational tools or first-pass information retrieval aids within ERAS pathways, but not as autonomous clinical decision-support systems. Clinician oversight, transparent reporting standards, and periodic re-evaluation as model versions evolve are essential preconditions for any clinical integration.

Limitations include only two clinician raters, a single query per question, and a single query date; model outputs, interfaces, and version labels may change over time. Model versions were not locked via API, meaning the underlying model weights could differ from those evaluated. The question bank, while comprehensive, may not fully capture the complexity of real-world multifactorial clinical decisions. The scoring system involves inherent subjectivity, reinforcing the importance of transparent and reproducible evaluation frameworks[4].

5. Conclusions

Both models showed high concordance with ERAS colorectal recommendations in a structured, blinded evaluation. Nevertheless, safety-relevant deviations and variability in rater agreement highlight the need for clinician oversight and standardized evaluation prior to clinical adoption.

Author Contributions: Conceptualization, study design, data curation, analysis, and manuscript drafting: S.G.

Acknowledgements: The author thanks the independent clinician raters (an anesthesiologist and a general surgeon) for their evaluations.

Generative AI Disclosure: Generative AI tools were used to assist with language editing and formatting under the author's supervision. The evaluated LLMs (Google Gemini and OpenAI ChatGPT) were used only to generate the study outputs and were not used to fabricate data; all analyses, interpretations, and final decisions were performed by the author.

Ethical Approval: Not applicable. This study did not involve human participants or patient-level data.

Conflict of Interest: The author declares no conflicts of interest.

Funding: No external funding was received.

Data Availability: The question bank, prompt template, and scoring rubric are provided as supplementary materials. De-identified rating data are available from the author upon reasonable request.

References

1. Gustafsson UO, Rockall TA, Wexner S, et al. Guidelines for perioperative care in elective colorectal surgery: Enhanced Recovery After Surgery (ERAS) Society recommendations 2025. Surgery. 2025;184:109397. doi:10.1016/j.surg.2025.109397

2. Gustafsson UO, et al. Guidelines for Perioperative Care in Elective Colorectal Surgery: Enhanced Recovery After Surgery (ERAS®) Society Recommendations: 2018. World J Surg. 2019;43(3):659-695. doi:10.1007/s00268-018-4844-y

3. Hager P, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1

4. Gallifant J, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5

5. Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
6. Palenzuela DL, Mullen JT, Phitayakorn R. AI Versus MD: Evaluating the surgical decision-making accuracy of ChatGPT-4. Surgery. 2024;176(2):241-245. doi:10.1016/j.surg.2024.04.003
7. Chatziisaak D, et al. Concordance of ChatGPT artificial intelligence decision-making in colorectal cancer multidisciplinary meetings: retrospective study. BJS Open. 2025;9(3):zraf040. doi:10.1093/bjsopen/zraf040
8. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213-220. doi:10.1037/h0026256
9. Schwarzkopf S-C, Bereuter J-P, Geissler ME, et al. Postoperative complication management: How do large language models measure up to human expertise? PLOS Digit Health. 2025;4(8):e0000933. doi:10.1371/journal.pdig.0000933
10. Kwon DY, et al. Adherence of a Large Language Model to Clinical Guidelines for Craniofacial Plastic and Reconstructive Surgeries. Ann Plast Surg. 2024;92(3):261-262. doi:10.1097/SAP.0000000000003757