Check for updates

Clinical Anatomy

WILEY

CLINICAL ANATOMY

EDUCATION

Large Language Models and Applied Surgical Anatomy: Evaluation of Anatomical Reasoning in Expert-Derived Plastic Surgery Scenarios

Menekşe Kastamoni Başkan[1] | Süleyman Çeçen[1] | Yadigar Kastamoni[2]

[1]Department of Plastic, Reconstructive and Aesthetic Surgery, Uludag University Faculty of Medicine, Bursa, Turkey | [2]Department of Anatomy, Suleyman Demirel University Faculty of Medicine, Isparta, Turkey

Correspondence: Menekşe Kastamoni Başkan (mkastamoni@uludag.edu.tr)

Received: 16 June 2026 | **Revised:** 25 September 2026 | **Accepted:** 29 September 2026

Keywords: anatomical reasoning | clinical decision-making | critical error analysis | large language models | plastic surgery | surgical anatomy

ABSTRACT

Large language models (LLMs) are increasingly being integrated into medical education, research, and clinical decision-support applications; however, their ability to apply anatomical knowledge in clinically relevant surgical scenarios remains insufficiently studied. This study evaluated and compared the performance of six contemporary LLMs (ChatGPT, Gemini, Claude, DeepSeek, Copilot, and Manus) using 25 expert-derived applied surgical anatomy scenarios relevant to plastic surgery. Scenarios were developed by an expert panel consisting of one associate professor of anatomy and two plastic surgeons. Responses were generated using a standardized prompt and assessed with a structured scoring rubric evaluating anatomical accuracy, clinical reasoning, completeness, and safety, with a maximum score of 10 points per question. Critical errors were additionally assessed. Significant differences in overall performance were observed among models (Friedman $\chi^2 = 43.51$, $p < 0.001$). ChatGPT achieved the maximum aggregate score (250/250), followed by Gemini (248/250), DeepSeek (241/250), Copilot (240/250), Claude (235/250), and Manus (219/250). Critical errors were infrequent and were confined to a reconstructive vascular physiology scenario: DeepSeek, Claude, and Manus each demonstrated one observed critical error, whereas ChatGPT, Gemini, and Copilot demonstrated none. Although all evaluated models demonstrated excellent performance, clinically significant anatomical misconceptions persisted despite high aggregate scores. These findings demonstrate strong performance by the tested LLMs on this specific scenario-based benchmark, while also identifying clinically meaningful anatomical errors in a subset of responses. Future evaluations should focus not only on overall accuracy but also on clinically meaningful error patterns that may influence surgical decision-making. Importantly, substantial ceiling effects limited discrimination among high-performing models, and the findings should therefore be interpreted as performance on this specific benchmark rather than as evidence of a general hierarchy of anatomical reasoning capability.

1 | Introduction

Large language models (LLMs) are increasingly being explored across medical education, research, clinical communication, and decision support (Abi-Rafeh et al. 2024; Song and Xu 2024; Snee et al. 2024; Mete and Özmen 2025; Hack, Alsleibi, et al. 2025; Meyer et al. 2025). Plastic surgery has similarly investigated their use in scientific writing, literature synthesis, patient communication, clinical documentation, and educational applications (Abi-Rafeh et al. 2024; Najafali, Hinson, et al. 2023; Najafali, Camacho, et al. 2023; Lim et al. 2024). Although these systems demonstrate substantial potential, concerns regarding factual accuracy, reproducibility, transparency, fabricated information, and patient safety continue to require

© 2026 American Association of Clinical Anatomists and British Association of Clinical Anatomists.

All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies.

expert oversight and validation (Abi-Rafeh et al. 2024; Najafali, Hinson, et al. 2023; Lim et al. 2024).

Several studies have evaluated LLM performance using plastic surgery educational and examination-based assessments. DiDonna et al. compared seven AI platforms using the Plastic Surgery In-Service Training Examination (PSITE) and demonstrated superior performance of ChatGPT-4 compared with competing models, although substantial variability persisted among platforms (DiDonna et al. 2024). Najafali et al. similarly evaluated GPT-4 and GPT-3.5 using the Plastic Surgery In-Service Examination and found that GPT-4 outperformed its predecessor while still falling short of senior resident-level performance in several comparisons (Najafali et al. 2025). These studies suggest a growing role for LLMs as educational adjuncts in surgical training. However, they primarily assessed examination-style knowledge rather than clinically applied anatomical reasoning.

Beyond examination-based assessments, investigators have explored the role of LLMs in plastic surgery decision-making. Borna et al. evaluated ChatGPT-4 and Gemini in emergency plastic surgery decision-making scenarios and found that ChatGPT-4 generally outperformed Gemini in diagnosis and management, while emphasizing that AI should complement rather than replace clinical evaluation and expertise (Borna et al. 2024). Gomez-Cabello et al. assessed ChatGPT-4 and Gemini for intraoperative decision support in plastic surgery and reported that ChatGPT-4 provided more accurate and relevant responses, although both models showed performance inconsistency across procedures (Gomez-Cabello et al. 2024). In rhinoplasty, Hack et al. evaluated multiple LLMs against professional guideline-based questions and found generally high accuracy and reliability, while still emphasizing the need for clinical oversight (Hack, Gvili, et al. 2025). Collectively, these studies demonstrate growing interest in LLM-assisted surgical reasoning, but they do not specifically address clinically applied surgical anatomy.

Anatomy remains one of the foundational sciences underpinning all surgical specialties and is particularly critical in plastic and reconstructive surgery. A precise understanding of nerves, vessels, muscles, fascial planes, perforators, and three-dimensional anatomical relationships directly influences operative planning, intraoperative decision-making, complication avoidance, and patient outcomes. Even minor anatomical misconceptions may lead to significant surgical morbidity. Consequently, the ability of LLMs to accurately apply anatomical knowledge represents a clinically important area of investigation. Despite the central role of anatomy in surgical practice, relatively few studies have specifically examined how accurately modern LLMs apply anatomical knowledge in clinically relevant contexts.

Growing interest in AI-assisted anatomy education has resulted in multiple studies evaluating the educational potential of conversational AI systems. Abdellatif et al. emphasized the continuing importance of anatomy as a cornerstone of medical education and discussed the potential role of AI in anatomy teaching, learning, and assessment (Abdellatif et al. 2022). Li et al. developed an AI-powered anatomy chatbot using a machine learning architecture and highlighted the potential of dialogue-based AI systems for anatomy education (Li et al. 2021). Totlis et al. evaluated ChatGPT-4 through an anatomy-focused conversation and found that it could provide structured anatomical explanations, while recognizing limitations in anatomical variants and clinical applicability (Totlis et al. 2023). Kundakcı also highlighted the potential of ChatGPT in anatomy education and anatomy research, while noting that errors may become more likely as topics become more complex (Kundakcı 2023). However, these studies largely focused on educational applications, factual anatomical knowledge, student engagement, or anatomy teaching tools. Relatively little attention has been directed toward evaluating the ability of modern LLMs to apply anatomical knowledge in clinically relevant surgical scenarios.

Importantly, existing studies predominantly assess anatomy through factual recall questions, educational interventions, or examination-based performance metrics. To date, no study has systematically evaluated multiple frontier LLMs using expert-derived applied surgical anatomy scenarios specific to plastic surgery while simultaneously examining clinically meaningful anatomical error patterns. Furthermore, direct comparisons among several contemporary LLMs using standardized expert-derived anatomy scenarios remain scarce.

Therefore, the aim of the present study was to evaluate and compare the performance of six contemporary LLMs in clinically applied surgical anatomy relevant to plastic surgery. Using 25 expert-derived clinically relevant anatomical scenarios and a structured scoring system assessing anatomical accuracy, clinical reasoning, completeness, and safety, we sought to characterize current model capabilities, identify clinically meaningful error patterns, and determine whether significant differences exist among modern LLMs in their ability to apply anatomical knowledge to clinically contextualized plastic surgery scenarios.

2 | Materials and Methods

2.1 | Study Design

This study was designed as a comparative evaluation of contemporary LLMs using expert-derived applied surgical anatomy scenarios relevant to plastic surgery practice. The primary objective was to compare model performance in anatomically complex clinical situations requiring integration of anatomical knowledge, surgical reasoning, and complication recognition. The study did not involve human participants, patient data, or identifiable clinical information.

2.2 | Scenario Development

A question bank consisting of 25 clinically relevant surgical anatomy scenarios was developed collaboratively by one associate professor of anatomy and two plastic surgeons with academic and clinical experience in reconstructive surgery. Scenarios were specifically designed to reflect anatomically relevant decision-making situations encountered in plastic surgery practice.

The scenarios were organized into five predefined domains:

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1. Facial Nerve and Facial Surgery (Questions 1–5)
2. Orbital and Periorbital Surgery (Questions 6–10)
3. Hand Surgery (Questions 11–15)
4. Breast and Abdominal Reconstruction (Questions 16–20)
5. Head and Neck and Lower Extremity Reconstruction (Questions 21–25)

All scenarios were designed to require application of anatomical knowledge rather than simple factual recall. Questions focused on identification of critical anatomical structures, interpretation of clinically relevant anatomy, operative decision-making, complication recognition, and anatomy-based surgical reasoning.

The complete expert-derived question bank is provided in Appendix S1 to facilitate transparency, reproducibility, and future comparative evaluations of LLMs in applied surgical anatomy.

2.3 | LLMs Evaluated

Six publicly available contemporary LLM platforms were evaluated: ChatGPT (OpenAI; GPT-5.5 Instant), Gemini (Google; Gemini 3.5 Flash), Claude (Anthropic; Claude Sonnet 4.6), DeepSeek (DeepSeek AI; DeepSeek V4), Microsoft Copilot, and Manus (Manus AI; Manus 1.6). All models were accessed on June 13, 2026 through their respective official web interfaces using a desktop web browser. Only publicly available free-access tiers were used; no paid, premium, or professional subscription was used for any model. For platforms in which the web interface did not explicitly disclose the underlying foundation model or dynamically routed requests among models, no specific underlying model was inferred retrospectively.

Detailed technical characteristics and access conditions for each evaluated platform are provided in Table S4.

2.4 | Response Generation

A standardized prompt was used for all models and all questions:

> You are an experienced academic plastic surgeon with expertise in surgical anatomy. Please answer the following question. Structure your response under the following headings:
>
> 1. Most Likely Anatomical Structure Involved
>
> 2. Anatomical Rationale
>
> 3. Clinical Significance
>
> Requirements:
>
> - Use concise academic language.
> - Focus on applied surgical anatomy relevant to plastic surgery.
> - Do not provide references.
> - Maximum response length: 150 words.
>
> Question:

To minimize carryover effects and contextual influence from previous interactions, each question was submitted in a new independent chat session. All models received identical prompts and identical question wording. Follow-up questions, clarification requests, regeneration of responses, and additional prompting were not permitted. Only the initial response generated by each model was included in the analysis. Each model–scenario combination was queried once; therefore, the study evaluated a single sampled response from each model for each scenario rather than repeated-response performance. The complete unedited model-generated responses for all 150 model–scenario combinations are provided in Appendix S3.

All platforms were used under their default free-tier web-interface configurations. No optional tools or ancillary features were manually activated or modified by the investigators. Specifically, web search/browsing, extended or deep-reasoning modes, file upload, and other optional tools were not intentionally invoked during response generation. The study sessions were initiated without prior study-related conversational context, and each scenario was submitted in a separate new chat. Because some platforms may implement proprietary background routing or platform-level features that are not visible to users, the internal activation status of undisclosed features could not be independently verified. Descriptive observations regarding response language, adherence to the requested format, and unsolicited follow-up behavior are provided in Table S2.

2.5 | Reference Answer Key Development

A preliminary answer key was developed by the study investigators for all 25 scenarios. The answer key was subsequently reviewed and refined by the same expert panel involved in scenario development, consisting of one associate professor of anatomy and two plastic surgeons. No additional independent expert panel was used to externally validate the scenarios or reference answer key. For each scenario, the expert panel established a reference standard consisting of:

- Essential anatomical elements
- Expected clinical reasoning components
- Key clinical implications
- Potential dangerous misconceptions or errors

This expert-derived answer key served as the reference standard for all scoring procedures.

2.6 | Scoring Framework

Model responses were evaluated using a predefined scoring rubric with a maximum score of 10 points per question. Scoring

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

was based on a predefined four-domain framework assessing Anatomical Accuracy, Clinical Reasoning, Completeness, and Safety. Evaluation was guided by the expected anatomical content for each scenario, including the key anatomical structure or concept, relevant anatomical and clinical reasoning, acceptable alternative formulations, and potentially misleading or unsafe interpretations.

Scoring was performed by two evaluators with complementary expertise, an associate professor of anatomy and a plastic surgeon, neither of whom was involved in generating the LLM responses. Before evaluation, model-identifying information was removed from the responses, and both evaluators were blinded to model identity throughout the scoring process. The two evaluators reviewed each response jointly with reference to the predefined scenario-specific answer key and assigned a single consensus score for each scoring domain. Any differences in interpretation were discussed until a single agreed score was reached. Thus, the study used blinded consensus-based scoring rather than independent parallel ratings. Because separate rater-level scores were not generated, formal inter-rater reliability statistics could not be calculated.

The rubric consisted of four domains:

- Anatomical Accuracy (0–4 points)

Assessment of correct identification and description of relevant anatomical structures and relationships.

- Clinical Reasoning (0–3 points)

Assessment of logical interpretation of anatomical findings and integration into clinical decision-making.

- Completeness (0–2 points)

Assessment of inclusion of all essential components expected in the reference answer.

- Safety (0–1 point)

Assessment of whether the response contained an explicitly unsafe, harmful, or contraindicated clinical recommendation. A score of 1 was assigned when no such recommendation was present, whereas a score of 0 was assigned when the response explicitly advocated an action that, if followed, could plausibly result in patient harm. An incorrect anatomical or physiological explanation alone did not result in loss of the safety point unless it was accompanied by an explicitly unsafe clinical recommendation.

Operational anchors for each score level within the four scoring domains are provided in Table S5.

The maximum possible score for each response was 10 points, resulting in a maximum cumulative score of 250 points per model (25 questions × 10 points).

The complete scoring dataset, including question-level scores, subscores, and critical-error annotations for all evaluated responses, is provided in Appendix S2.

2.7 | Critical Error Assessment

In addition to numerical scoring, responses were evaluated for critical errors using the same blinded consensus-based procedure.

A critical error was defined as a clearly incorrect anatomical or physiological interpretation with plausible potential to materially misdirect surgical decision-making or result in an unsafe clinical inference. Mere deviation from the preferred reference answer, omission of a nonessential element, or selection of an alternative interpretation that was anatomically or clinically defensible was not considered a critical error.

Critical-error classification and the Safety subscore were operationally distinct. A critical error reflected a materially incorrect anatomical or physiological interpretation with potential to misdirect clinical reasoning, whereas loss of the Safety point required an explicitly harmful, unsafe, or contraindicated clinical recommendation. Accordingly, a response could contain a critical error while retaining the Safety point if the response contained a clinically consequential misconception but did not explicitly recommend an unsafe action.

Critical-error classification was recorded separately from the numerical scoring system and analyzed descriptively.

During manuscript revision, disputed critical-error classifications were independently reassessed by an additional plastic surgeon who had not participated in scenario development, reference-answer development, initial scoring, or LLM response generation. The reassessment was performed with model identity concealed, and affected scores and statistical analyses were updated accordingly.

2.8 | Statistical Analysis

Statistical analyses were performed using question-level scores as the unit of analysis. Continuous variables were summarized as mean ± standard deviation, median, and range where appropriate.

Differences in overall performance among models were assessed using the Friedman test. Post hoc pairwise comparisons were performed using Wilcoxon signed-rank tests with Holm correction for multiple comparisons. Because the paired score differences contained numerous zero differences and ties, Wilcoxon signed-rank p values were calculated using the asymptotic normal approximation rather than an exact procedure. Effect size was estimated using Kendall's coefficient of concordance (Kendall's W).

Domain-specific analyses were performed separately for each predefined anatomical domain using the Friedman test, with five matched questions contributing to each domain. Because of the small number of questions per domain and the presence of extensive ties and ceiling effects, domain-level inferential results were interpreted cautiously and considered together with descriptive score distributions and Kendall's W effect sizes. Friedman test p values were based on the asymptotic

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

chi-square approximation. To account for multiplicity across domain-level omnibus comparisons, Holm adjustment was applied to the four domains for which a Friedman test could be performed. The Hand Surgery domain showed no score variation across models and therefore did not permit a meaningful omnibus test. Critical-error frequencies were summarized descriptively.

A two-sided p value <0.05 was considered statistically significant. Because the 25 scenarios constituted a purposively constructed benchmark rather than a random sample from a defined population of plastic surgery anatomy questions, statistical inference was restricted to between-model differences within this specific scenario set and was not intended to provide population-level estimates for plastic surgery anatomy as a whole.

3 | Results

3.1 | Overall Model Performance

A total of 150 responses generated by six LLMs (ChatGPT, Gemini, DeepSeek, Copilot, Claude, and Manus) across 25 expert-derived surgical anatomy scenarios were evaluated using the predefined scoring rubric (maximum score: 10 points per question; total possible score: 250 points per model). Significant differences in overall performance were observed among the evaluated models (Friedman $\chi^2 = 43.51$, df = 5, $p < 0.001$, Kendall's $W = 0.348$), indicating a moderate effect size.

Within the sampled response set, ChatGPT achieved a total score of 250/250 (mean score 10.00 ± 0.00), followed by Gemini (248/250, 9.92 ± 0.28), DeepSeek (241/250, 9.64 ± 1.25), Copilot (240/250, 9.60 ± 1.22), Claude (235/250, 9.40 ± 1.66), and Manus (219/250, 8.76 ± 1.81). Median scores were 10 for all models except Manus, which demonstrated a median score of 9. Detailed overall performance metrics are presented in Table 1.

Post hoc pairwise Wilcoxon signed-rank testing with Holm correction demonstrated significantly lower performance for Manus compared with ChatGPT (adjusted $p = 0.012$), Gemini (adjusted $p = 0.015$), Claude (adjusted $p = 0.012$), and Copilot (adjusted $p = 0.012$). No statistically significant differences were identified among the remaining model pairs following correction for multiple comparisons (Table S1).

The distribution of overall model performance is illustrated in Figure 1.

3.2 | Domain-Specific Performance

Domain-specific analyses demonstrated heterogeneous performance patterns across the evaluated anatomical domains. ChatGPT achieved perfect scores across all five domains (50/50 in each domain), whereas Gemini demonstrated near-perfect performance with minor score reductions in the Facial Nerve and Facial Surgery and Breast and Abdominal Reconstruction domains.

TABLE 1 | Overall performance metrics and observed critical-error frequency across evaluated models.

Model	*n*	Total/250	Percent	Mean	SD	Median	IQR	Min	Max	Full-score responses	Critical errors	Observed critical-error frequency
ChatGPT	25	250	100.00	10.00	0.00	10	10–10	10	10	25	0	0.0%
Gemini	25	248	99.20	9.92	0.28	10	10–10	9	10	23	0	0.0%
DeepSeek	25	241	96.40	9.64	1.25	10	10–10	4	10	22	1	4.0%
Copilot	25	240	96.00	9.60	1.22	10	10–10	4	10	20	0	0.0%
Claude	25	235	94.00	9.40	1.66	10	10–10	4	10	20	1	4.0%
Manus	25	219	87.60	8.76	1.81	9	8–10	3	10	11	1	4.0%

Note: Observed critical-error frequency was calculated as the number of sampled responses containing at least one critical error divided by the 25 responses generated by each model. Because each model–scenario combination was sampled once, these percentages describe the observed response set and should not be interpreted as estimates of model-specific error probabilities across repeated generations.

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

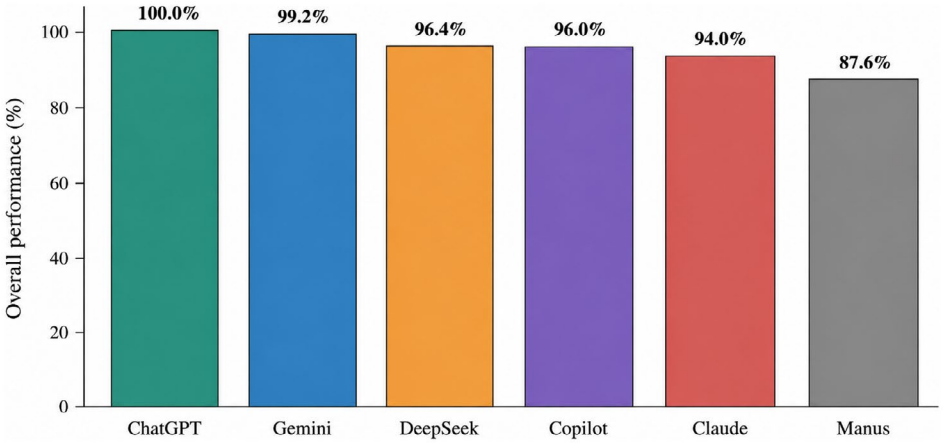


FIGURE 1 | Overall performance of the evaluated large language models across 25 expert-derived applied surgical anatomy scenarios. Performance is expressed as the percentage of the maximum achievable aggregate score (250 points) for each model. ChatGPT achieved 100.0%, followed by Gemini (99.2%), DeepSeek (96.4%), Copilot (96.0%), Claude (94.0%), and Manus (87.6%).

TABLE 2 | Domain-specific total scores.

Model	Facial nerve and facial surgery	Orbital and periorbital surgery	Hand surgery	Breast and abdominal reconstruction	Head and neck/ lower extremity reconstruction
ChatGPT	50	50	50	50	50
Gemini	49	50	50	49	50
DeepSeek	50	48	50	49	44
Copilot	49	44	50	48	49
Claude	50	44	50	47	44
Manus	45	43	50	41	40

Note: Maximum score per domain = 50. Domain-level inferential analyses were based on five matched questions per domain. Holm-adjusted p values and Kendall's W were as follows: Facial Nerve and Facial Surgery, $p = 0.031$, $W = 0.600$; Orbital and Periorbital Surgery, $p = 0.700$, $W = 0.120$; Breast and Abdominal Reconstruction, $p = 0.010$, $W = 0.733$; Head and Neck/Lower Extremity Reconstruction, $p = 0.046$, $W = 0.522$. The Hand Surgery domain showed no between-model variation and was not subjected to an omnibus significance test.

Manus consistently achieved the lowest domain-specific scores, particularly in Breast and Abdominal Reconstruction (41/50) and Head and Neck/Lower Extremity Reconstruction (40/50). Detailed domain-specific results are presented in Table 2.

After Holm adjustment across the four testable domain-level omnibus comparisons, between-model differences remained statistically significant in Facial Nerve and Facial Surgery (Friedman $\chi^2 = 15.00$, Holm-adjusted $p = 0.031$, Kendall's $W = 0.600$), Breast and Abdominal Reconstruction ($\chi^2 = 18.33$, Holm-adjusted $p = 0.010$, Kendall's $W = 0.733$), and Head and Neck/Lower Extremity Reconstruction ($\chi^2 = 13.05$, Holm-adjusted $p = 0.046$, Kendall's $W = 0.522$). Orbital and Periorbital Surgery did not demonstrate a statistically significant between-model difference ($\chi^2 = 3.00$, Holm-adjusted $p = 0.700$, Kendall's $W = 0.120$). Hand Surgery demonstrated complete ceiling-level performance across all evaluated models, with no between-model score variation; therefore, a meaningful Friedman omnibus test could not be performed for this domain. Given the small number of questions within each domain and the extensive ties, these inferential findings should be interpreted together with the descriptive domain scores and effect sizes presented in Table 2 and Figure 2.

3.3 | Critical Error Analysis

Critical errors were infrequent across the sampled responses. ChatGPT, Gemini, and Copilot demonstrated no critical errors, whereas DeepSeek, Claude, and Manus each demonstrated one critical error (4.0%). These percentages represent observed frequencies within the single sampled response set and should not be interpreted as estimates of model-specific critical-error probabilities across repeated generations. A summary of observed critical-error frequencies is provided in Table 3.

The remaining critical-error pattern occurred in Question 25 (Head and Neck/Lower Extremity Reconstruction domain). Claude, DeepSeek, and Manus interpreted the scenario primarily as a venous outflow or congestion problem rather than identifying the angiosome–perforasome/choke-vessel mechanism underlying impaired distal perfusion despite preserved arterial inflow. This pattern was retained as a critical error because the responses attributed the finding to a different physiological mechanism rather than the vascular-territory mechanism specified by the scenario and reference framework. These responses nevertheless retained the Safety point because they did

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

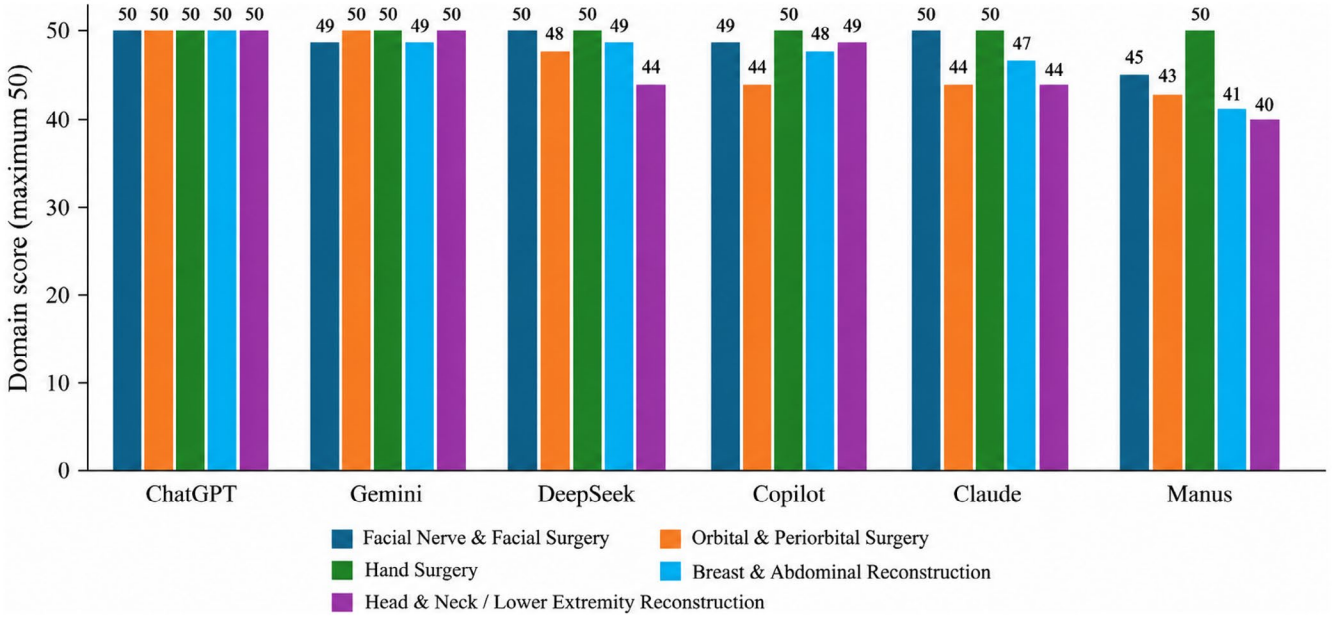


FIGURE 2 | Domain-specific performance of the evaluated large language models. Scores are presented for the five predefined anatomical domains: Facial Nerve and Facial Surgery, Orbital and Periorbital Surgery, Hand Surgery, Breast and Abdominal Reconstruction, and Head and Neck/Lower Extremity Reconstruction. Maximum score per domain was 50 points.

TABLE 3 | Observed critical-error frequency by model.

Model	Critical errors	Observed critical-error frequency
ChatGPT	0	0.0%
Gemini	0	0.0%
DeepSeek	1	4.0%
Copilot	0	0.0%
Claude	1	4.0%
Manus	1	4.0%

Note: Observed critical-error frequency represents the proportion of the 25 sampled responses from each model that contained at least one critical error; it should not be interpreted as a stable model-specific error rate across repeated generations.

not explicitly recommend a harmful or contraindicated clinical action; the critical-error designation reflected the materially incorrect physiological interpretation rather than an explicitly unsafe management recommendation. The critical-error pattern is summarized in Table 4.

3.4 | Subscore Analysis

Subscore analysis demonstrated consistently high performance across all evaluated models. ChatGPT achieved maximum scores in all evaluation domains, including Anatomical Accuracy (100/100), Clinical Reasoning (75/75), Completeness (50/50), and Safety (25/25). Gemini demonstrated near-maximal performance with only minor reductions in Anatomical Accuracy and Clinical Reasoning. Manus demonstrated the lowest performance in Anatomical Accuracy (88/100) and Clinical Reasoning (58/75), while Completeness and Safety scores remained relatively preserved across all models. Detailed subscore analyses are presented in Table S3.

Overall, although all evaluated models demonstrated high levels of performance, the overall omnibus comparison indicated significant between-model differences, and a clinically meaningful critical-error pattern persisted in the reconstructive vascular physiology scenario despite generally excellent total scores.

4 | Discussion

In this study, we evaluated six contemporary LLMs using 25 expert-derived applied surgical anatomy scenarios relevant to plastic surgery. Within this single sampled response set, aggregate scores were high across all evaluated models, with ChatGPT obtaining 250/250 and Gemini 248/250. These findings characterize performance on the sampled responses generated under the study conditions and should not be interpreted as estimates of stable or intrinsic model performance. Nevertheless, statistically significant differences were observed among models, and a clinically meaningful critical-error pattern remained identifiable despite generally excellent aggregate scores. These findings indicate that the tested LLMs performed strongly on this specific scenario-based assessment while remaining vulnerable to selected anatomical misconceptions with potential clinical relevance.

The overall performance observed in the present study is consistent with the rapidly expanding literature demonstrating increasing LLM capabilities across medical education, surgical training, and specialty-specific knowledge assessments. Previous investigations have reported that newer-generation

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Critical-error pattern identified in Question 25.

Question	Domain	Models with critical error	Error pattern	No. models affected
Q25	Head and Neck/ Lower Extremity Reconstruction	Claude; DeepSeek; Manus	Venous outflow/congestion interpretation instead of choke vessel/ angiosome-perforasome concept	3

Note: Critical-error classification was reserved for interpretations judged to represent a clearly incorrect anatomical or physiological interpretation with plausible potential to materially misdirect surgical decision-making or result in an unsafe clinical inference.

models frequently achieve strong performance on specialty-specific educational and examination-based assessments, including those used in plastic surgery training (DiDonna et al. 2024; Najafali et al. 2025). Our findings extend these observations by demonstrating that high performance is not limited to factual recall or examination-style questions but also applies to clinically contextualized anatomy scenarios requiring integration of anatomical knowledge and surgical reasoning.

Recent studies evaluating LLMs in plastic surgery decision-making have likewise reported encouraging results. Borna et al. demonstrated strong performance of ChatGPT and Gemini in emergency plastic surgery scenarios, while Gomez-Cabello et al. reported favorable results in intraoperative decision-support tasks (Borna et al. 2024; Gomez-Cabello et al. 2024). Hack et al. similarly found that LLMs provided generally accurate and reliable responses to rhinoplasty guideline-based questions, but concluded that clinical oversight remained essential for safe use in patient education and decision-making (Hack, Gvili, et al. 2025). The present findings are consistent with the broader observation that high aggregate performance does not preclude clinically meaningful errors. Although overall scores were exceptionally high, anatomically significant errors continued to occur in a subset of scenarios, highlighting the distinction between global performance metrics and clinically meaningful reliability.

One notable finding was that the sampled ChatGPT responses received the maximum aggregate score of 250/250 across the 25 scenarios. This finding should not be interpreted as evidence of error-free or deterministic model performance. Because each model–scenario combination was sampled only once, alternative responses generated in independent sessions could differ owing to the stochastic nature of LLM output generation. Accordingly, the observed score represents the performance of the response set generated under the specific study conditions rather than an estimate of the model's stable underlying capability. The question bank was intentionally designed to reflect core anatomical concepts encountered in plastic surgery practice rather than obscure anatomical trivia. However, high performance in controlled scenario-based assessments does not necessarily imply equivalent performance in real-world clinical environments, where patient-specific variability, incomplete information, visual data, operative nuance, and contextual uncertainty are common.

The presence of a ceiling effect represents an important consideration when interpreting our findings. Several domains, particularly Hand Surgery, demonstrated near-perfect performance across all evaluated models. This ceiling-level performance reduces the discriminatory capacity of the benchmark, particularly among high-performing models. Nevertheless, the assessment of clinically relevant anatomical error patterns provides complementary information beyond aggregate score differences. From this perspective, the persistence of critical errors despite excellent overall scores may be more informative than modest differences in aggregate performance. In other words, a model achieving 95–100% accuracy may still generate a clinically consequential anatomical misconception capable of influencing surgical judgment.

This concern is especially relevant in surgical anatomy. Abdellatif et al. emphasized that anatomical knowledge remains fundamental to clinical examination, diagnosis, treatment planning, and surgical or invasive procedures (Abdellatif et al. 2022). Totlis et al. similarly highlighted that anatomical knowledge is required for understanding disease location, performing procedures safely, interpreting imaging, and communicating precisely among healthcare professionals (Totlis et al. 2023). In plastic surgery, where operative decisions frequently depend on fine anatomical distinctions involving nerves, vessels, muscles, fascial planes, perforators, and flap territories, a single misconception may have direct procedural consequences. Therefore, the evaluation of LLMs in this field should not rely only on general correctness but should also examine whether models avoid clinically meaningful anatomical errors.

The reassessment of disputed responses also highlighted the importance of distinguishing a critical error from a defensible alternative interpretation. In Question 7, the inferior oblique muscle represented the preferred reference answer for vertical diplopia following lower blepharoplasty; however, published clinical evidence indicates that the inferior rectus muscle may also be injured during lower blepharoplasty and may contribute to postoperative diplopia (Galli 2012; Becker 2020). Accordingly, responses identifying the inferior rectus were no longer classified as critical errors following reassessment, although they remained below the preferred reference answer within the numerical scoring framework. In contrast, the critical-error designation was retained for Question 25, in which three models attributed impaired distal flap perfusion primarily to venous outflow dysfunction rather than the angiosome–perforasome/ choke-vessel mechanism targeted by the scenario (Taylor et al. 2017; Saint-Cyr et al. 2009). This distinction illustrates the importance of reserving critical-error classification for interpretations that exceed simple deviation from the preferred reference answer and represent a materially different anatomical or physiological mechanism.

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The analysis of critical errors may represent the most clinically relevant contribution of the present study. Many previous LLM evaluations have focused primarily on overall accuracy, examination scores, or response quality (DiDonna et al. 2024; Najafali et al. 2025; Borna et al. 2024; Gomez-Cabello et al. 2024; Hack, Gvili, et al. 2025). While these metrics are valuable, they may obscure anatomically significant misconceptions embedded within otherwise convincing responses. Our findings demonstrate that modern LLMs can simultaneously achieve excellent overall performance while still generating potentially consequential errors in reconstructive vascular physiology. Consequently, future evaluations should incorporate carefully defined critical-error assessment in addition to traditional accuracy-based metrics, while distinguishing unequivocally incorrect or clinically consequential interpretations from defensible alternatives to the preferred reference answer.

The domain-specific analyses provide additional insight into model performance. Significant differences were observed in Facial Nerve and Facial Surgery, Breast and Abdominal Reconstruction, and Head and Neck/Lower Extremity Reconstruction domains, whereas Hand Surgery demonstrated ceiling-level performance across all models. These findings should be interpreted cautiously given the small number of scenarios within each domain, extensive ties, and the purposive construction of the benchmark. Accordingly, the observed differences characterize performance on this specific scenario set rather than establishing a general hierarchy of anatomical reasoning capability. Future benchmark iterations should incorporate more challenging and heterogeneous scenarios involving anatomical variation, ambiguous operative findings, complex spatial relationships, imaging-based anatomy, competing anatomical explanations, and multistep anatomy-based decision-making to improve discrimination and construct validity.

This study has several limitations. First, the 25 scenarios were purposively developed by the study expert panel and were not intended to represent the broader population of plastic surgery anatomy questions. Moreover, the same panel developed and reviewed the reference answer key without independent external validation, introducing potential circularity and limiting content validity and generalizability. Second, each model–scenario combination was queried only once. Because LLM outputs are stochastic, the observed scores and critical-error frequencies characterize a single sampled response set and cannot establish within-model reproducibility or stable model performance. Retrospective repetition under identical conditions was not feasible because the model versions, routing mechanisms, and free-tier platform configurations available on June 13, 2026 could not be reliably recreated after subsequent platform updates. Third, the study therefore represents publicly available free-access model configurations at a single time point and should not be assumed to characterize earlier or subsequent versions. Fourth, responses were scored jointly by two blinded evaluators using consensus rather than independent parallel ratings; consequently, formal inter-rater reliability could not be calculated. Fifth, substantial ceiling effects, particularly complete ceiling-level performance in the Hand Surgery domain, limited discrimination among high-performing models and constrained the comparative validity of small aggregate-score differences. Finally, the study evaluated text-based responses to simulated clinical scenarios rather than real-world performance and did not assess learner or patient outcomes, multimodal interpretation, or actual clinical decision support. Accordingly, the findings should be interpreted as performance on this specific benchmark rather than evidence of broader educational or clinical utility.

Despite these limitations, the study possesses several important strengths. The question bank was specifically developed by experts in anatomy and plastic surgery to incorporate clinically relevant and anatomically focused scenarios. Unlike many previous studies that relied primarily on examination questions, educational prompts, or factual knowledge assessments, the present investigation utilized clinically contextualized scenario-based evaluations requiring integration of anatomical knowledge, clinical reasoning, and complication recognition. This design enabled assessment of clinically meaningful anatomical error patterns that may not be detected through traditional accuracy-based evaluation methods. Additionally, the inclusion of six contemporary LLMs, a structured expert-derived reference framework, domain-specific analyses, and critical-error assessment provides a comprehensive framework for future investigations in surgical anatomy and medical artificial intelligence.

5 | Conclusions

The sampled responses from contemporary LLMs demonstrated generally high performance on this expert-derived benchmark of applied surgical anatomy, with relatively few critical errors observed in the sampled response set. However, the substantial ceiling effect, including complete ceiling-level performance in the Hand Surgery domain, limited discrimination among high-performing models. Accordingly, these findings should be interpreted as performance on this specific benchmark rather than as evidence of a general hierarchy of anatomical reasoning capability. Future benchmarks incorporating greater anatomical complexity, ambiguity, spatial reasoning, and multimodal information are needed to better characterize the limits of LLM performance in applied surgical anatomy.

Funding

The authors have nothing to report.

Ethics Statement

The authors have nothing to report.

Consent

The authors have nothing to report.

Data Availability Statement

The data supporting the findings of this study are provided in the Supporting Information accompanying this article.

References

Abdellatif, H., M. Al Mushaiqri, H. Albalushi, A. A. Al-Zaabi, S. Roychoudhury, and S. Das. 2022. "Teaching, Learning and Assessing

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Anatomy With Artificial Intelligence: The Road to a Better Future." *International Journal of Environmental Research and Public Health* 19, no. 21: 14209. https://doi.org/10.3390/ijerph192114209.

Abi-Rafeh, J., H. H. Xu, R. Kazan, R. Tevlin, and H. Furnas. 2024. "Large Language Models and Artificial Intelligence: A Primer for Plastic Surgeons on the Demonstrated and Potential Applications, Promises, and Limitations of ChatGPT." *Aesthetic Surgery Journal* 44, no. 3: 329–343. https://doi.org/10.1093/asj/sjad260.

Becker, B. B. 2020. "Diplopia Following Lower Blepharoplasty." *Journal of AAPOS* 24, no. 6: 363.e1–e4. https://doi.org/10.1016/j.jaapos.2020.07.017.

Borna, S., C. A. Gomez-Cabello, S. M. Pressman, S. A. Haider, and A. J. Forte. 2024. "Comparative Analysis of Large Language Models in Emergency Plastic Surgery Decision-Making: The Role of Physical Exam Data." *Journal of Personalized Medicine* 14, no. 6: 612. https://doi.org/10.3390/jpm14060612.

DiDonna, N., P. N. Shetty, K. Khan, and L. Damitz. 2024. "Unveiling the Potential of AI in Plastic Surgery Education: A Comparative Study of Leading AI Platforms' Performance on In-Training Examinations." *Plastic and Reconstructive Surgery. Global Open* 12, no. 6: e5929. https://doi.org/10.1097/GOX.0000000000005929.

Galli, M. 2012. "Diplopia Following Cosmetic Surgery." *American Orthoptic Journal* 62: 19–21. https://doi.org/10.3368/aoj.62.1.19.

Gomez-Cabello, C. A., S. Borna, S. M. Pressman, S. A. Haider, and A. J. Forte. 2024. "Large Language Models for Intraoperative Decision Support in Plastic Surgery: A Comparison Between ChatGPT-4 and Gemini." *Medicina (Kaunas, Lithuania)* 60, no. 6: 957. https://doi.org/10.3390/medicina60060957.

Hack, S., S. Alsleibi, N. Saleh, E. E. Alon, N. Rabinovics, and E. Remer. 2025. "Are Chatbots a Reliable Source for Patient Frequently Asked Questions on Neck Masses?" *European Archives of Oto-Rhino-Laryngology* 282, no. 8: 4273–4282. https://doi.org/10.1007/s00405-025-09433-6.

Hack, S., B. Gvili, S. Alsleibi, D. Yogev, and E. Glikson. 2025. "Do Large Language Models Meet Professional Standards in Rhinoplasty? A Comparative Evaluation With AAO-HNS Guidelines." *European Journal of Plastic Surgery* 48: 71. https://doi.org/10.1007/s00238-025-02329-y.

Kundakcı, Y. E. 2023. "ChatGPT's Capabilities for Use in Anatomy Education and Anatomy Research." *European Journal of Therapeutics* 30, no. 2: 200–202. https://doi.org/10.58600/eurjther1842.

Li, Y. S., C. S. N. Lam, and C. See. 2021. "Using a Machine Learning Architecture to Create an AI-Powered Chatbot for Anatomy Education." *Medical Science Educator* 31, no. 6: 1729–1730. https://doi.org/10.1007/s40670-021-01405-9.

Lim, B., I. Seth, Y. Xie, P. S. Kenney, R. Cuomo, and W. M. Rozen. 2024. "Exploring the Unknown: Evaluating ChatGPT's Performance in Uncovering Novel Aspects of Plastic Surgery and Identifying Areas for Future Innovation." *Aesthetic Plastic Surgery* 48, no. 13: 2580–2589. https://doi.org/10.1007/s00266-024-03952-z.

Mete, U., and Ö. A. Özmen. 2025. "Assessing the Accuracy and Reproducibility of ChatGPT for Responding to Patient Inquiries About Otosclerosis." *European Archives of Oto-Rhino-Laryngology* 282: 1567–1575. https://doi.org/10.1007/s00405-024-09039-4.

Meyer, A., Y. Karay, A. U. Steinbicker, T. Streichert, and R. Overbeek. 2025. "Performance of DeepSeek-R1, ChatGPT (GPT-o3-Mini), and Gemini 2.0 Flash on German Medical Multiple-Choice Questions: Comparative Evaluation." *JMIR Formative Research* 9: e77357. https://doi.org/10.2196/77357.

Najafali, D., J. M. Camacho, E. Reiche, L. G. Galbraith, S. D. Morrison, and A. H. Dorafshar. 2023. "Truth or Lies? The Pitfalls and Limitations of ChatGPT in Systematic Review Creation." *Aesthetic Surgery Journal* 43, no. 8: NP654–NP655. https://doi.org/10.1093/asj/sjad093.

Najafali, D., C. Hinson, J. M. Camacho, L. G. Galbraith, R. Gupta, and C. M. Reid. 2023. "Can Chatbots Assist With Grant Writing in Plastic Surgery? Utilizing ChatGPT to Start an R01 Grant." *Aesthetic Surgery Journal* 43, no. 8: NP663–NP665. https://doi.org/10.1093/asj/sjad116.

Najafali, D., E. Reiche, S. Araya, et al. 2025. "Artificial Intelligence Augmentation: Performance of GPT-4 and GPT-3.5 on the Plastic Surgery In-Service Examination." *Plastic and Reconstructive Surgery. Global Open* 13, no. 4: e6645. https://doi.org/10.1097/GOX.0000000000006645.

Saint-Cyr, M., C. Wong, M. Schaverien, A. Mojallal, and R. J. Rohrich. 2009. "The Perforasome Theory: Vascular Anatomy and Clinical Implications." *Plastic and Reconstructive Surgery* 124, no. 5: 1529–1544. https://doi.org/10.1097/PRS.0b013e3181b98a6c.

Snee, I., C. X. Lava, K. R. Li, and G. D. Corral. 2024. "The Utility of ChatGPT in Gender-Affirming Mastectomy Education." *Journal of Plastic, Reconstructive & Aesthetic Surgery* 99: 432–435. https://doi.org/10.1016/j.bjps.2024.10.020.

Song, Y., and T. Xu. 2024. "Letter to the Editor for the Article 'Performance of ChatGPT-3.5 and ChatGPT-4 on the European Board of Urology (EBU) Exams: A Comparative Analysis'." *World Journal of Urology* 42: 555. https://doi.org/10.1007/s00345-024-05256-y.

Taylor, G. I., R. J. Corlett, and M. W. Ashton. 2017. "The Functional Angiosome: Clinical Implications of the Anatomical Concept." *Plastic and Reconstructive Surgery* 140, no. 4: 721–733. https://doi.org/10.1097/PRS.0000000000003694.

Totlis, T., K. Natsis, D. Filos, et al. 2023. "The Potential Role of ChatGPT and Artificial Intelligence in Anatomy Education: A Conversation With ChatGPT." *Surgical and Radiologic Anatomy* 45, no. 10: 1321–1329. https://doi.org/10.1007/s00276-023-03229-1.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** ca70243-sup-0001-AppendixS1.docx. **Appendix S2:** ca70243-sup-0002-AppendixS2.xlsx. **Appendix S3:** ca70243-sup-0003-AppendixS3.docx. **Table S1:** Pairwise post hoc Wilcoxon signed-rank tests (Holm corrected). **Table S2:** Response behavior/instruction adherence. **Table S3:** Subscore totals by model. **Table S4:** Technical characteristics and access conditions of the evaluated large language model platforms. **Table S5:** Operational scoring anchors for the four-domain response evaluation rubric.

10982353, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ca.70243 by Meneke Kastamoni Bakan - Bursa Uludag University Kutuphane ve Dokumantasyon , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License