

REVIEW ARTICLE

Generative AI in Surgical Care: Evaluating Large Language Model Performance in Patient Education

Anne E. Hall¹, Archi K. Patel¹, Kaavian Shariati¹, Amanda T. Perrotta¹, Alexander A. Argame², Charles Y. Tseng¹, Chi-Hong Tseng³, Marco A. Hidalgo^{3,4}, Justine C. Lee^{1,4,*}

¹Division of Plastic and Reconstructive Surgery, Department of Surgery, University of California Los Angeles, David Geffen School of Medicine and the Greater Los Angeles VA Healthcare System, Los Angeles, CA, USA

²Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, USA

³Division of General Internal Medicine and Health Services Research, Department of Medicine, University of California Los Angeles, David Geffen School of Medicine, Los Angeles, CA USA

⁴UCLA Gender Health Program, University of California Los Angeles, David Geffen School of Medicine, Los Angeles, CA, USA

ARTICLE DATA

Article History

Received 28 July 2025

Revised 17 November 2025

Accepted 02 December 2025

Keywords

Artificial intelligence
Surgery
Large-language models
Patient questions
Accuracy
Quality
Readability

ABSTRACT

Background: The advancement of generative artificial intelligence (AI) technologies is affecting how individuals seek and consume information. Large language models (LLMs) are increasingly being used for health-related inquiries, including surgical information, yet concerns remain regarding the quality of evidence-based content they provide.

Methods: A systematic review per PRISMA guidelines was conducted to evaluate the accuracy, quality, and readability of LLMs in responding to surgery-related questions. Studies were retrieved from MEDLINE, CENTRAL, and EMBASE databases related to LLM-generated responses to patient-centered surgical inquiries (December 2024) and were included if they presented patient-centered surgical questions to an LLM, defined relevant outcome measures, and used quantitative evaluation scales to assess LLM-responses. Quantitative synthesis involved a linear mixed-effects model, using estimated marginal means and Holm-adjusted post-hoc comparisons to assess model performance.

Results: 64 studies comprising 1,664 surgical questions across 12 surgical subspecialties were included. The most evaluated LLMs were ChatGPT-3.5 (70.3%) followed by ChatGPT-4.0 (37.5%) and Google Gemini (23.4%). ChatGPT (pooled analysis of versions 3.0, 3.5 and 4.0) scored higher in accuracy compared to other LLMs ($74.7\% \pm 2.5$ vs $65.9\% \pm 3.2$, $p = 0.004$). However, all LLMs underperformed in quality compared to standard patient education materials, as measured by the DISCERN instrument. ChatGPT scored significantly lower (51.8 ± 1.8 vs 65.6 ± 2.4 , $p < 0.0001$), as did other LLMs (49.7 ± 2.7 vs 65.6 ± 2.4 , $p < 0.0001$), compared to standard patient education materials. Readability, measured by the Flesch-Kincaid Grade Level was higher in ChatGPT compared to other LLMs (13.4 ± 0.3 vs 10.5 ± 0.5 , $p < 0.0001$).

Conclusions: LLMs are an emerging tool in surgical patient education, but demonstrate limitations related to content quality. Further refinements in AI-generated health content can further improve clarity, reliability, and surgery-specific accuracy.

1. INTRODUCTION

As artificial intelligence (AI) continues to advance and become more accessible, its adoption across diverse fields has accelerated [1]. In particular, generative AI and large language models (LLMs) are increasingly being integrated into roles that require greater nuance and complexity [2]. In healthcare, providers and patients alike are more frequently turning to these tools for tasks ranging from medical education and

clinical decision support to administrative workflow improvement, patient engagement, and information retrieval [3–9]. Additionally, AI-driven chatbots, particularly those utilizing natural language processing (NLP) and machine learning, have gained traction for their ability to act as accessible, real-time sources of health information [10–12]. Their ability to enhance healthcare efficiency and personalize patient interactions has positioned them as promising tools in the evolving landscape of digital health technology [13–15].

*Corresponding author. Email: justine@ucla.edu

© 2025 The Authors. Published by Nexus Press B.V.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Among these emerging technologies, AI-driven chatbots such as the Chat Generative Pre-Trained Transformer (ChatGPT), have garnered global attention for their potential role in delivering healthcare information [16]. With advanced NLP capabilities, LLMs have the ability to bridge communication gaps and enhance efficiency by delivering real-time responses to health-related queries [15,17,18]. As patients increasingly seek medical information online, these tools present a unique opportunity to supplement traditional sources of healthcare information. In surgical contexts, both patients and providers are exploring the utility of LLMs. Patients may seek procedure-related information, guidance, and advice through LLMs, while healthcare providers may leverage them to streamline the delivery of educational materials, improving patient preparedness, perioperative support, and knowledge [15,19]. Thus far, the utility of ChatGPT has been assessed in various surgical specialties, including plastic and reconstructive surgery [20–22], orthopedic surgery [23], and vascular surgery [24,25]. Throughout the perioperative period, LLMs have been reported to assist with consultations [26,27], informed consent [28–30], and patient education, allowing individuals to ask questions related to their procedure [31–33]. Additionally, these models can provide personalized care [34,35], postoperative support [36–38], and guidance in response to complications [39–41].

While LLMs are increasingly regarded as a valuable tool for patients, their adoption in the broader surgical field, where risk is high and informed decision-making is critical, is still nascent. Integration has been somewhat less rapid in comparison to other fields like primary care and telemedicine, where LLMs are being applied for patient interactions and virtual consultations [42,43]. Despite the increasing pool of data that highlights the relevance of AI-driven chatbots to healthcare, concerns remain regarding the accuracy, usability, safety, and quality of evidence-based content provided by LLMs in the context of perioperative care [44–46]. Physicians and patients alike may have concerns as to whether the information provided is comprehensive and adequate, and the fidelity with which they view AI guidance and its influence on clinical decision-making varies [10,47,48]. Though LLMs can access a vast amount of data and approaches clinician-level performance and knowledge, further refinement may be required [49–51].

As interest in the applications of generative AI in healthcare continues to grow, a systematic evaluation of its use in providing surgical patient education is valuable. This review seeks to systematically assess the accuracy, quality, and readability of LLMs in responding to patient-centered surgical questions. Through an analysis of these performance metrics, this work aims to evaluate the effectiveness of AI-driven chatbots in providing accurate, reliable, high-quality surgical information.

2. METHODS

2.1. Search Criteria

A systematic search of MEDLINE, EMBASE, and Cochrane Central Register of Controlled Trials (CENTRAL) databases was conducted in December 2024 following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. To identify studies evaluating LLM performance metrics, the search strategy included terms such as “chatbot,” “surgery,” “accur*,” “readability,” “questions,” “patient,” and “education” (Supplementary File 1).

2.2. Inclusion and Exclusion Criteria

Article screening was performed by AEH and AKP. Inclusion criteria included studies evaluating LLMs in the context of surgical patient education and reported quantitative outcomes on various performance metrics including accuracy, quality, and readability using validated scales or scoring systems. Studies that did not assess LLM evaluation metrics using a detailed scoring methodology or did not report quantitative outcomes related to the quality of LLM-generated surgical patient education were excluded. The reference lists of all included full-text articles, as well as other relevant articles identified during the screening process, were examined to locate additional studies.

2.3. Data Extraction

Study characteristics were independently extracted by AEH and AKP, including surgical procedure, evaluation method, LLMs evaluated, non-LLM comparisons, number of evaluators, and outcomes of interest. For studies evaluating LLM accuracy, quality, and readability, additional extracted characteristics included evaluation metrics, study results, consensus-based or independent scoring, and the inter-rater agreement. Quality outcomes were assessed using various standardized metrics including the DISCERN instrument, JAMA Benchmark Criteria, Ensuring Quality Information for Patients (EQIP) and Global Quality Scores (GQS). The DISCERN instrument evaluates the reliability and quality of written health information, while the JAMA Benchmark Criteria assess source transparency based on authorship, attribution, currency, and disclosure [52,53]. Additionally, EQIP measures clarity and completeness in patient education materials, and GQS provides a global rating of information usefulness [54,55]. As for readability, the Flesch-Kincaid Grade Level (FKGL) and the Flesch Reading Ease Score (FRES) scales were most commonly used and extracted. FKGL estimates the U.S. grade level required to understand a text, with higher scores indicating greater reading difficulty [56]. FRES evaluates readability on a 0–100 scale, where higher scores correspond to easier-to-read content [57]. Any data not reported in the original studies are labeled as ‘not available (NA)’ in Table 1.

2.4. Statistical Analysis

Accuracy outcomes were separated into two groups: 1) those that reported an overall average score of the LLM responses and 2) those that reported the proportion of responses that achieved the highest possible score on their respective evaluation scale. Many studies employed Likert scales to assess accuracy; to standardize these measures, all scales were adjusted to ensure a consistent directionality (e.g., higher scores indicating greater accuracy). To facilitate comparison across studies, the average score for each study was converted into a percentage to normalize across different scoring scales (Supplementary File 2). For studies that did not report a standard deviation, a proxy value was imputed using the median standard deviation from the other included studies in the same outcome grouping.

As for quality and readability outcomes, due to the scoring differences among quality assessment tools (DISCERN, JAMA, EQIP, and GQS) and readability instruments (FKGL and FRES), we did not perform standardized mean difference (SMD) calculations or other cross-metric comparisons. Differences between detailed LLM classifications as well as broad LLM classifications

were calculated using a weighted linear mixed-effects model, with the study as the random intercept and the model as the fixed effect. The model was weighted by the number of questions and evaluations per study. Estimated marginal means and Holm-adjusted post-hoc comparisons was further used to assess model performance. An omnibus likelihood ratio test was conducted to evaluate the overall effect of model type on the outcome. In the broad LLM classifications, ChatGPT referred to ChatGPT-3.0, ChatGPT-3.5 and ChatGPT-4.0, whereas 'Other LLMs' included Google Gemini, Bing-Co-Pilot, Claude, Perplexity, and CLOVA X. Among evaluation of the broad LLM classifications, ChatGPT was set as the reference group. For detailed LLM classifications, LLMs assessed included: ChatGPT-3.0, ChatGPT-3.5, ChatGPT-4.0 and Google Gemini, while 'Other LLMs' encompassed Bing Co-Pilot, Claude, Perplexity, and CLOVA X. Additionally, MediSearch and OpenEvidence were considered as 'health LLMs' as they are specifically trained on medical literature, clinical guidelines, and scientific publications to improve the accuracy of medical queries. Among the evaluation of detailed LLM classifications, ChatGPT-3.5 was set as the reference group. Google search response scores were not included in the linear mixed model. All statistical analysis was conducted in R Studio (version 4.4.2, Vienna, Austria) using the 'lme4' and 'emmeans' packages, with $p < 0.05$ considered statistically significant.

3. RESULTS

3.1. Study Characteristics

A search of MEDLINE, EMBASE, and CENTRAL databases resulted in 295 unique articles for chatbot performance metrics, of which 108 met criteria for full-text review (Supplementary File 3). 42 of the 108 articles were subsequently excluded during review due to lack of reporting on quantitative outcomes ($n = 23$) or study design ($n = 19$). Inter-rater agreement for abstract screening and full-text review was 93.0% and was resolved through discussion. After abstract and full-text review screening, 64 articles were included in the review [20–25,44,58–114].

The included studies, published between 2023 and 2024, evaluating chatbot performance metrics encompassed a total of 1,664 surgical questions asked to LLMs across 50 unique procedures (Table 1). There was a total of 375 evaluators with 92.5% being surgeons from their respective fields. Most studies assessed LLM performance by presenting the question to the model, with an average of 26 questions per study (range: 3–151) and six evaluators (range: 1–45) conducting the assessments. Most evaluations (79.7%) were conducted independently, while 20.3% of studies required evaluators to reach a consensus before assigning a final score.

The studies primarily focused on a wide range of orthopedic surgeries, accounting for 27 out of the 64 included studies (42.2%). Other specialties analyzed included plastic and reconstructive surgery (7.8%), ophthalmology (7.8%), neurosurgery (7.8%), general surgery (7.8%), and oral and maxillofacial (4.7%). In terms of study outcomes, accuracy was the most assessed metric (57.8%), while 51.7% of studies evaluated readability and 37.5% evaluated the quality of LLM-generated responses.

Regarding the LLMs evaluated, the most frequently assessed models were ChatGPT-3.5 (70.3%) and ChatGPT-4.0 (37.5%), followed by Google Gemini (22.0%), Bing Co-Pilot (6.3%), and Claude (4.7%). Most studies evaluated a single LLM (67.2%),

while 20.3% evaluated two LLMs, and 12.5% evaluated three or more LLMs. Importantly, 20.3% of studies included a non-LLM comparator, with standard educational materials (53.8%) being the most common, followed by Google search responses (30.8%), and health care providers responses (15.4%).

3.2. Accuracy

Of the 37 studies evaluating accuracy, 73.0% calculated the overall average score of the responses based on a Likert scale, while 27.0% reported the proportion of responses that reached the highest Likert scale category (Table 2, Supplementary File 4). When converting the average score of the responses into a percentage, ChatGPT scored higher in accuracy compared to other LLMs ($74.7\% \pm 2.5$ vs $65.9\% \pm 3.2$, $p = 0.004$) and performed comparably to standard patient education materials, with no statistically significant difference (Table 3). Similarly, when examining detailed models and the proportion of responses that reached the highest Likert scale category, ChatGPT-4.0 outperformed Google Gemini ($75.5\% \pm 8.1$ vs $40.8\% \pm 7.5$, $p = 0.002$).

3.3. Quality

Of the 24 studies assessing quality, 58.3% utilized DISCERN, 25% utilized JAMA Benchmark Criteria, 12.5% utilized EQIP, and 12.5% utilized GQS (Table 3). Among studies that employed the DISCERN instrument, standard patient education materials scored higher than LLM-generated responses (65.6 ± 2.4 vs 51.8 ± 1.8 , $p < 0.0001$ for ChatGPT and 49.7 ± 2.7 for other AI models, $p < 0.0001$, Table 4). When analyzed by detailed model type, standard health materials outperformed each individual LLM (Table 5). Similarly, when applying the JAMA Benchmark Criteria, standard materials exceeded ChatGPT (2.6 ± 0.6 vs 0.3 ± 0.4 , $p = 0.002$), with differences observed across ChatGPT model versions (Table 5). For GQS, health LLMs also surpassed both ChatGPT and other LLMs (4.3 ± 0.2 vs 4.7 ± 0.2 , $p < 0.0001$ for ChatGPT and 4.1 ± 0.2 for other AI models, $p < 0.0001$, Table 4), with significant differences also observed between individual ChatGPT-3.5 and 4.0 versions (Table 5). As for the EQIP assessment, health LLMs outperformed ChatGPT (36.1 ± 6.8 vs 30.5 ± 6.7 , $p = 0.0002$); particularly in ChatGPT-4.0 responses (Table 5).

3.4. Readability

Of the 33 studies that assessed readability, 87.8% utilized FKGL and 66.7% utilized FRES (Table 3). Overall, ChatGPT required a higher educational level compared to other LLMs (13.4 ± 0.3 vs 10.5 ± 0.5 , $p < 0.0001$) as well as standard education materials (9.7 ± 0.6 , $p < 0.0001$). This trend was also observed further among individual ChatGPT models (Table 5). As for ease of readability, standard education materials had a greater ease of readability compared to ChatGPT (54.8 ± 3.2 vs 33.6 ± 1.5 , $p < 0.0001$) and other LLMs (47.0 ± 2.5 , $p = 0.04$). Among detailed LLM classifications, standard education materials outperformed ease of readability for each individual ChatGPT model (Table 5).

4. DISCUSSION

Despite their recent emergence, LLMs have been increasingly implemented across various surgical disciplines, where their

Table 1 | LLM performance study characteristics. This table summarizes key study characteristics including surgical procedure, evaluation method, large language models (LLMs) evaluated, non-LLM comparisons, evaluators, outcomes of interest and metrics used

Study	Surgical procedure	Evaluation method	LLMs evaluated	Non-LLM comparisons	Evaluators	Outcomes	Outcome metrics
Adelstein et al., 2024 [58]	Total shoulder arthroplasty	10 FAQs	ChatGPT-3.5	Evidence-based literature	2 orthopedic surgeons	Accuracy	Likert scale
Alessandri-Bonetti et al., 2023 [20]	Body contouring surgery	15 FAQs	ChatGPT-3.5	Google search	4 plastic surgeons	Quality	GQS
Al-Sharif et al., 2024 [59]	Oculoplastic surgeries	112 FAQs	ChatGPT-3.5, Google Gemini	None	4 oculoplastic specialists	Accuracy Empathy Readability	Likert scale Likert scale ARI
Al-Shehri et al., 2025 [60]	Hip arthroscopy	10 FAQs	ChatGPT-3.5	None	2 fellowship-trained hip preservation surgeons	Accuracy	Likert scale
Anastasio et al., 2023 [23]	Foot and ankle surgeries	5 FAQs	ChatGPT-3.5	None	11 orthopedic surgeons and fellows	Readability Quality Educational value	FKGL, FRES DISCERN AIM
Bains et al., 2024 [61]	Total knee arthroplasty	60 FAQs	ChatGPT-4.0	Responses from arthroplasty-trained nurses	3 fellowship-trained surgeons	Appropriateness Patient comfort	Likert scale Binary scale and Likert scale
Balel 2023 [21]	Oral and maxillofacial surgery	60 FAQs	ChatGPT-3.5	None	33 surgeons	Quality	Modified GQS
Carlson et al., 2024 [62]	Vasectomy	15 FAQs	ChatGPT-3.5, Google Gemini, Bing Co-pilot, Perplexity, Claude	None	2 attending urologists, 2 urology research faculty, 1 urological resident physician	Accuracy Readability	Likert scale FKGL, FRES
Casciato et al., 2024 [63]	Foot and ankle surgeries	20 FAQs	ChatGPT-3.5	FootCareMD	2 podiatrists	Readability Quality	FKGL, FRES JAMA benchmark criteria, DISCERN
Chervonski et al., 2024 [25]	Vascular surgeries	24 FAQs	ChatGPT-3.5, Google Gemini	None	6 vascular surgeons	Accuracy Completeness Appropriateness Readability	Likert scale Likert scale Binary scale FKGL, FRES, GFI
Cohen et al., 2024 [64]	Cataract surgery	20 FAQs	ChatGPT-3.5	Google search	3 board-certified ophthalmologists and 1 ophthalmology senior resident	Accuracy Safety Readability	Likert scale Likert scale FKGL, FRES, GFI, SMOG, CLI
Crook et al., 2023 [65]	Hand surgery	67 FAQs	ChatGPT-3.0	ASSH HandCare materials	NA	Quality Readability	JAMA benchmark criteria, DISCERN FKGL, FRES

(Continued)

Table 1 | (Continued)

Study	Surgical procedure	Evaluation method	LLMs evaluated	Non-LLM comparisons	Evaluators	Outcomes	Outcome metrics
Endo et al., 2023 [66]	Liver transplant surgery	29 FAQs	ChatGPT-3.5	None	17 experts in the field of abdominal transplant surgery	Accuracy	Likert scale
Eng et al., 2025 [67]	Rotator cuff repair	10 FAQs	ChatGPT-3.5	None	2 orthopedic surgeons	Reliability Quality Readability	JAMA benchmark DISCERN FKGL
Fahy et al., 2024 [68]	ACL surgery	33 FAQs	ChatGPT-3.5, 4.0	None	3 orthopedic surgeons specializing in knee surgery	Quality Readability	DISCERN FKGL, FRES, SMOG, GFI
Ferrari-Light et al., 2025 [69]	Lung cancer surgery	30 FAQ	ChatGPT-3.5	None	9 thoracic surgeons	Quality	Likert scale
Frodl et al., 2024 [70]	Patellofemoral surgery	12 FAQs	ChatGPT-4.0	None	4 experts in patellofemoral disorders	Accuracy Quality	Likert scale EQIP
Gajjar et al., 2024 [71]	Neurosurgical procedures	60 FAQs	ChatGPT-3.0, 3.5, 4.0	None	5 attending neurosurgeons and 5 neurosurgery nurses	Helpfulness Accuracy Comprehensiveness Actionability Readability	Likert scale Likert scale Likert scale Likert scale FRES, SMOG, CLI, GFI, RRE
Garg et al., 2024 [22]	Aesthetic facial plastic surgeries	28 FAQs	ChatGPT-3.5, Google Gemini	None	1 attending surgeon, 1 FPS fellow, and 1 senior resident	Accuracy Readability	QOS FKGL
Grippaudo et al., 2024 [72]	Breast surgeries	3 FAQs	ChatGPT-3.5	None	10 plastic surgery residents	Quality	EQIP
Haidar et al., 2023 [24]	Vascular surgeries	3 FAQs	ChatGPT-3.5	Circular Foundation UK	2 surgical residents	Readability	FKGL, SMOG, GFI, FRES
Hu et al., 2024 [73]	Orthopedic surgeries (PJI post-THA and TKA)	12 FAQs	ChatGPT-4.0	None	1 senior infectiologist, 1 senior orthopedic surgeon, 1 resident infectiologist	Quality Accuracy	DISCERN Likert scale
Hurley et al., 2024 [74]	Shoulder stabilizing surgery	23 FAQs	ChatGPT-3.5	None	3 orthopedic residents	Quality	JAMA benchmark criteria, DISCERN
Ichhpujani et al., 2024 [75]	Glaucoma surgery	25 FAQs	ChatGPT-3.5, Google Gemini	None	2 glaucoma specialists	Readability Appropriateness Readability	FKGL, FRES Likert scale FKGL, FRES, GFI, CLI, ARI, SMOG

Jo et al., 2024 [76]	Colorectal cancer surgery	10 FAQs	ChatGPT-4.0, Google Gemini, CLOVA X	CRC information book	5 healthcare professionals	Accuracy Completeness Empathy Patient satisfaction Humanness of answer	Likert scale Likert scale Likert scale Likert scale Likert scale
Johns et al., 2024 [77]	ACL surgery	10 FAQs	ChatGPT-3.5	None	3 orthopedic residents	Readability	FKGL, SMOG, CLI, GFI, ARI
Johns et al., 2024 [78]	UCL reconstruction	10 FAQs	ChatGPT-3.5	None	2 orthopedic residents and 1 medical student	Quality	DISCERN
						Accuracy	Likert scale
Kaarre et al., 2023 [79]	ACL surgery	20 FAQs	ChatGPT-4.0	None	4 sport medicine orthopedic surgeons	Accuracy	Likert scale
						Readability	FKGL
Kaarre et al., 2024 [80]	Orthopedic trauma surgery	34 FAQs	ChatGPT-4.0	None	3 orthopedic trauma surgeons	Accuracy	Likert scale
						Completeness	Likert scale
Kasapovic et al., 2024 [81]	Orthopedic and trauma surgeries	20 LLM-generated information sheets	ChatGPT-3.5	None	2 orthopedic surgeons	Adaptability	Likert scale
						Quality	Modified DISCERN
Lack et al., 2024 [82]	Reverse shoulder arthroplasty	10 FAQs	ChatGPT-3.5	None	2 orthopedic surgeons	Reliability	Global medical assessment
						Quality	JAMA benchmark criteria, DISCERN
Lang et al., 2024 [83]	Common lumbar spine fusion surgery	10 FAQs	ChatGPT-3.5, Gemini	None	5 spine surgeons	Readability	FKGL
						Accuracy	Likert scale
Lang et al., 2024 [84]	Disc herniation surgery	10 FAQs	ChatGPT-3.5, Gemini	None	6 spine surgeons	Comprehensiveness	Likert scale
						Empathy	Likert scale
Lauck et al., 2024 [85]	Microscopic mohs surgery	10 FAQs	ChatGPT-3.5, Gemini	Google search	15 MMS surgeons	Accuracy	Likert scale
						Empathy	Likert scale
Lim et al., 2024 [86]	Abdominoplasty	15 FAQs	ChatGPT-3.5, Gemini, Claude, Bing Co-Pilot	None	2 plastic surgery residents	Appropriateness	Binary scale
						Accuracy	Likert scale
						Sufficiency	Likert scale
						Readability	FRES, FKGL
						Readability	FKGL, FRES, CLI
						Quality	DISCERN, Likert scale

(Continued)

Table 1 | (Continued)

Study	Surgical procedure	Evaluation method	LLMs evaluated	Non-LLM comparisons	Evaluators	Outcomes	Outcome metrics
Lois et al., 2024 [87]	Inguinal hernia information	17 FAQs	ChatGPT-3.5, 4.0, Google Gemini, Claude	Two gold-standard educational materials provided by SAGES and ACS	6 hernia surgeons	Accuracy Relevance Completeness Readability Quality Quality	Likert scale Likert scale Likert scale FKGL, FRES DISCERN JAMA benchmark criteria
Mastrokostas et al., 2024 [88]	Anterior cervical discectomy and fusion	10 FAQs	ChatGPT-4.0	Google search	2 independent reviewers	Readability Accuracy	FRES, FKGL Likert scale
Mika et al., 2023 [89]	Total hip arthroplasty	10 FAQs	ChatGPT-3.5	None	11 adult joint reconstruction fellowship-trained surgeons	Accuracy	Likert scale
Mika et al., 2024 [90]	Total knee arthroplasty	10 FAQs	ChatGPT-3.5	None	2 orthopedic surgeons	Accuracy	Likert scale
Momenaei et al., 2023 [91]	Retinal surgeries	88 FAQs	ChatGPT-4.0	None	2 retina specialists	Appropriateness Readability	Likert scale FKGL, FRES, GFI, CLI, SMOG
Mouhawasse et al., 2024 [92]	Vasectomy	10 FAQs	ChatGPT-3.5, Google Gemini, Bing Co-Pilot	None	3 urologists	Accuracy	Likert scale
Munir et al., 2024 [93]	GI surgical operations	24 FAQs	ChatGPT-4.0	None	45 general surgeons	Satisfaction Quality Utility	Likert scale Likert scale Likert scale
Nwokeji et al., 2024 [94]	Oral surgery	15 FAQs	ChatGPT-3.5, 4.0, Google Gemini, Bing Co-pilot	None	14 OMFS surgeons	Accuracy	Likert scale
Oliveria et al., 2017 [95]	Deep brain stimulation	40 FAQs	ChatGPT-3.5, 4.0	None	2 specialists in functional neurosurgery	Accuracy Completeness Danger	Likert scale Likert scale Likert scale
Patil et al., 2024 [96]	Ophthalmic surgeries	30 FAQs	ChatGPT-4.0, Google Gemini	None	2 ophthalmologists	Accuracy	Likert scale
Polat et al., 2024 [97]	Adenoidectomy, tonsillectomy, and ventilation tube insertion surgery	33 FAQs	ChatGPT-4.0	None	7 ENT surgeons	Accuracy Readability	Likert scale FKGL, FRES
Reddy et al., 2024 [98]	Vasectomy	20 FAQs	ChatGPT-3.0	None	3 fellowship-trained sexual medicine specialists	Accuracy	Likert scale
Robinson et al., 2024 [99]	Mohs surgery	12 FAQs	ChatGPT-4.0, Google Gemini	ACMS	11 Mohs surgeons	Accuracy Utility Readability	Likert scale Likert scale FKGL

Roldan- Vasquez et al., 2024 [44]	Breast cancer surgery	9 FAQs	ChatGPT-3.5	None	4 breast surgical oncologists	Accuracy Reliability	Likert scale Likert scale and PEMAT
Sahin et al., 2024 [100]	Thyroid surgery	21 FAQs	ChatGPT-4.0	None	3 ENT specialists	Quality Readability	DISCERN FKGL, FRES, GFI, SMOG, CLI, ARI
Sahin et al., 2024 [101]	Osteosclerosis surgery	15 FAQs	ChatGPT-4.0	None	3 ENT specialists	Quality Readability	DISCERN FKGL, FRES, GFI, SMOG, CLI, ARI
Samaan et al., 2023 [102]	Bariatric surgery	151 FAQs	ChatGPT-3.5	None	2 board-certified bariatric surgeons	Comprehensiveness Reproducibility	Likert scale Binary scale
Samsonov et al., 2024 [103]	Total ankle replacement	10 FAQs	ChatGPT-3.5	None	NA	Quality	Likert scale
Sarikonda et al., 2024 [104]	Minimally invasive spine surgery	15 FAQs	ChatGPT-3.5	None	3 neurosurgery attendings	Appropriateness Readability	Binary scale FKGL, FRES, SMOG, GFI, CLI, ARI
Shaari et al., 2024 [105]	UCL reconstruction	24 FAQs	ChatGPT-3.5, 4.0	None	2 orthopedic surgeons	Readability	FRES, FKGL, GFI, CLI, SMOG, ARI
Smith et al., 2024 [106]	Orthopedic surgeries	60 FAQs	ChatGPT-4.0	None	23 orthopedic surgeons	Accuracy Accuracy	Likert scale Likert scale
Snee et al., 2024 [107]	Gender-affirming mastectomy education	10 FAQs	ChatGPT-3.5	None	7 plastic surgeons specializing in GAM	Accuracy Comprehensiveness Danger Readability	Likert scale Likert scale Likert scale FKGL, FRES, GFI, CLI, SMOG, ARI
Soroudi et al., 2024 [108]	Breast reconstruction	10 FAQs	ChatGPT-3.0, 4.0	Responses from plastic surgeons and APP	2 plastic surgeons and 2 medical students	Accuracy Empathy Readability	Likert scale Likert scale FRES
Villarreal-Espinosa et al., 2024 [109]	ACL surgery	10 FAQs	ChatGPT-3.5	None	2 fellowship trained surgeons	Quality Reproducibility	Likert scale Binary scale
White et al., 2025 [110]	Total shoulder arthroplasty	15 FAQs	ChatGPT-3.5	None	5 orthopedic surgeons	Accuracy	Likert scale
Wright et al., 2024 [111]	Total hip arthroplasty and total knee arthroplasty	30 FAQs	ChatGPT-3.5	None	5 orthopedic residents	Accuracy Comprehensiveness Readability Acceptability	Likert scale Likert scale FKGL Binary scale

(Continued)

Table 1 | (Continued)

Study	Surgical procedure	Evaluation method	LLMs evaluated	Non-LLM comparisons	Evaluators	Outcomes	Outcome metrics
Yan et al., 2024 [112]	Cerebrovascular events in neurosurgical setting	13 FAQs	ChatGPT-3.5, 4.0	None	3 neurosurgeons	Readability Appropriateness	FKGL, GFI Binary scale
Yurdakurban et al., 2024 [113]	Orthognathic surgery	60 FAQs	ChatGPT-4.0, Medisearch, Open Evidence	None	2 independent reviewers	Quality Reliability	EQIP, GQS Reliability scoring system
Zhang et al., 2024 [114]	Total knee replacement	50 FAQs	ChatGPT-3.5	None	2 orthopedic surgeons	Readability Similarity Accuracy Relevance	SMOG Similarity index Likert scale Likert scale

ACL: Anterior Cruciate Ligament; ACS: American College of Surgeons; AIRM: Assessing the Impact of Research on Medicine; APP: advanced practice practitioners; ARI: Automated Readability Index; ASSH: American Society for Surgery of the Hand; CLI: Coleman-Liau Index; CRC: Colorectal Cancer; DISCERN: Discerning the Quality of Health Information; ENT: Ear, Nose, and Throat; EQIP: Ensuring Quality Information for Patients; FAQ: Frequently-Asked Question; FKGL: Flesch-Kincaid Grade Level; FRES: Flesch Reading Ease Score; FPS: Facial Plastic Surgery; GAM: Gender-Affirming Mastectomy; GFI: Gunning Fog Index; GQS: Global Quality Score; JAMA: Journal of the American Medical Association Benchmark Criteria; MMS: Mohs Micrographic Surgery; OMS: Oral Maxillofacial Surgery; PJI: Periprosthetic Joint Infection; RRE: Raygor Readability Estimate; SAGES: Society of American Gastrointestinal and Endoscopic Surgeons; SMOG: Simple Measure of Gobbledygook; TKA: Total Knee Arthroplasty; UCL: Ulmar Collateral Ligament.

effectiveness in addressing patient-centered surgical inquiries is the subject of ongoing investigation [115–117]. This systematic review evaluated the performance of LLMs in responding to patient-centered surgical questions, focusing on accuracy, quality, and readability. We found that LLM-generated responses were generally accurate when measured against standard patient education material benchmarks, with ChatGPT demonstrating the highest accuracy while operating at a college-level reading comprehension. Among the LLMs assessed with validated quality outcome metrics, all underperformed relative to standard patient education materials in their reliability, transparency, and adherence to evidence-based guidelines. However, quality outcomes varied between select models with health LLMs (MediSearch and OpenEvidence) achieving higher quality scores in comparison to ChatGPT-3.5 according to GQS and EQIP metrics. This suggests that LLMs specifically designed to accommodate health-related inquiries may potentially resolve suboptimal quality standards broadly met by these models.

Our results are consistent with the broader literature assessing LLM accuracy in answering patient-centered health inquiries. In our analysis, ChatGPT achieved a pool accuracy of 74.7% across surgical questions which aligns with a meta-analysis by Wei et al., who reported an overall accuracy of 56% across a broader range of medical specialties [118]. Additionally, in their cross-sectional study evaluating LLM accuracy when provided clinical oncology cases, Chen et al., found that ChatGPT-4.0 was among the highest performing LLMs according to free text evaluation, with 37.9% of responses deemed as ‘correct’ [119]. However, there has not yet been a published systematic review that specifically investigates the ability of LLMs to answer surgery-related patient questions. Instead, existing literature primarily explores the broader applications of LLMs in surgery, including their role in administrative tasks, surgical consent, operative report and discharge summary generation, surgical planning and research ideation [6,7,120,121]. The paucity of research in this area and the resulting inability to draw generalized, informed conclusions are particularly important given the expanding role of LLMs as a primary source of information for patients. While numerous studies have evaluated how appropriately ChatGPT and other LLMs answer surgery-related questions, a systematic evaluation of their performance is still lacking. Our study addresses these gaps by incorporating a variety of surgical procedures and systematically assessing multiple performance metrics, affording a comprehensive analysis of LLM effectiveness in the context of perioperative patient care. Additionally, we identified which models are most effective in delivering reliable and accessible surgical information, offering insights into their potential integration in patient education and clinical practice.

A notable challenge identified in this and other studies is suboptimal quality of responses provided by LLMs [122]. While the quality of patient questions has not yet undergone systematic evaluation in either surgical or large clinical settings, Wang et al., conducted a systematic review of the applications and concerns regarding conversational LLMs in healthcare, finding that 89% of included studies expressed concerns about reliability, bias, or both [123]. This finding is particularly significant given that LLMs gather data through large-scale web scraping, publicly available datasets, and reinforcement learning, and have been observed to generate hallucinations, a phenomenon where fabrication or misleading information appears credible but lacks factual bias [124]. In fact, in their comparative analysis assessing the performance of LLMs directed to provide

Table 2 | Likert scale accuracy results. This table presents the results of all studies that employed a Likert scale to assess accuracy. “Directionality” refers to whether a higher scale value indicates greater accuracy (high scale) or a lower scale value represents higher accuracy (low scale). The percent accuracy score was calculated by converting the average score into a percentage to normalize across different scoring scales

LLM and educational materials	Point Likert scale	Directionality	Likert score mean	SD	Percent accuracy score
<i>ChatGPT-3.5</i>					
Adelstein et al., 2024 [58]	4	LS	2.5	0.7	50.0%
Chervonski et al., 2024 [25]	4	HS	3.1	0.3	69.3%
Carlson et al., 2024 [62]	6	HS	5.0	0.7	80.8%
Gajjar et al., 2024 [71]	5	HS	4.4	0.7	85.0%
Johns et al., 2024 [77]	4	LS	3.2		27.3%
Johns et al., 2024 [78]	4	LS	2.8		40.7%
Lauck et al., 2024 [85]	5	HS	3.9		74.3%
Lois et al., 2024 [87]	6	HS	4.2	0.8	64.4%
Mika et al., 2023 [89]	4	LS	2.3	1.0	56.7%
Mika et al., 2024 [90]	4	LS	2.3	0.7	56.7%
Mouhawasse et al., 2024 [92]	4	LS	1.4	0.5	87.7%
Nwokeji et al., 2024 [94]	5	LS	2.1		71.5%
Roldan-Vasquez et al., 2024 [44]	5	HS	4.2		81.3%
Shaari et al., 2024 [105]	5	HS	3.6		65.8%
Snee et al., 2024 [107]	10	HS	8.8	0.5	86.7%
Villarreal-Espinosa et al., 2024 [109]	5	HS	4.4	1.0	85.0%
White et al., 2025 [110]	5	LS	2.9	1.0	53.0%
Wright et al., 2024 [111]	5	HS	4.3		81.3%
Zhang et al., 2024 [114]	5	HS	4.6		90.0%
<i>ChatGPT-4.0</i>					
Gajjar et al., 2024 [71]	5	HS	4.1	0.8	78.3%
Hu et al., 2024 [73]	4	LS	2.3	0.6	58.3%
Jo et al., 2024 [76]	5	HS	3.4	0.4	60.0%
Kaarre et al., 2023 [79]	3	HS	1.7		83.0%
Lois et al., 2024 [87]	5	HS	4.4		85.8%
Nwokeji et al., 2024 [94]	5	LS	2.1		73.8%
Patil et al., 2024 [96]	5	HS	4.5	0.6	87.5%
Robinson et al., 2024 [99]	5	HS	4.8	0.2	94.0%
Shaari et al., 2024 [105]	5	HS	3.9		73.0%
Smith et al., 2024 [106]	3	HS	2.4	0.6	71.5%
Soroudi et al., 2024 [108]	5	HS	4.7	0.5	92.5%
<i>ChatGPT-3.0</i>					
Gajjar et al., 2024 [71]	5	HS	3.9	1.0	73.3%
Reddy et al., 2024 [98]	5	HS	4.2	0.9	79.0%
Soroudi et al., 2024 [108]	5	HS	4.3	0.7	82.5%
<i>Google Gemini</i>					
Carlson et al., 2024 [62]	6	HS	4.4	1.1	68.6%
Chervonski et al., 2024 [25]	4	HS	2.8	0.4	60.7%
Jo et al., 2024 [76]	5	HS	2.9	0.4	47.5%
Lois et al., 2024 [87]	5	HS	4.1		76.5%
Mouhawasse et al., 2024 [92]	4	LS	2.2	0.9	61.0%
Nwokeji et al., 2024 [94]	5	LS	2.3		67.8%
Patil et al., 2024 [96]	5	HS	3.8	0.8	70.0%
Robinson et al., 2024 [99]	5	HS	4.1	0.4	76.8%
<i>Other LLMs</i>					
<i>Bing Co-Pilot</i>					
Carlson et al., 2024 [62]	6	HS	4.9	0.8	78.2%

(Continued)

Table 2 | (Continued)

LLM and educational materials	Point Likert scale	Directionality	Likert score mean	SD	Percent accuracy score
Mouhawasse et al., 2024 [92]	4	LS	1.8	0.6	73.3%
Nwokeji et al., 2024 [94]	5	LS	2.6	3.4	60.5%
<i>Claude</i>					
Carlson et al., 2024 [62]	6	HS	4.7	1.0	73.0%
Lois et al., 2024 [87]	5	HS	3.9		71.3%
<i>Perplexity</i>					
Carlson et al., 2024 [62]	6	HS	4.4	1.1	68.2%
<i>CLOVA X</i>					
Jo et al., 2024 [76]	5	HS	3.0	0.6	50.0%
<i>Standard educational materials and provider responses</i>					
Jo et al., 2024 [76]	5	HS	3.1	0.5	52.5%
Robinson et al., 2024 [99]	5	HS	4.5	0.3	87.5%
Soroudi et al., 2024 [108]	5	HS	4.2	0.9	80.0%
Soroudi et al., 2024 [108]	5	HS	4.4	0.9	85.0%
<i>Google search</i>					
Lauck et al., 2024 [85]	5	HS	3.6		64.8%

APP: Advanced Practice Practitioner; HS: High-Scale; LS: Low-Scale; SD: Standard Deviation.

Table 3 | Quality and readability results. This table presents the results of all studies that evaluated quality with the DISCERN instrument, JAMA benchmark criteria, the Ensuring Quality Information for Patients (EQIP) and the Global Quality Score (GQS), as well as all studies that evaluated readability utilizing the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES)

LLM and educational materials	Quality	Readability	
	Mean (SD)	FKGL Mean (SD)	FRES Mean (SD)
<i>ChatGPT-3.5</i>			
Alessandri-Bonetti et al., 2023 [20]	4.4 (0.7) ^a		
Al-Shehri et al., 2025 [60]		14.4 (1.8)	28.2 (9.2)
Anastasio et al., 2023 [23]	52.1 (18.5) ^b		
Balel 2023 [21]	4.6 (0.8) ^a		
Carlson et al., 2024 [62]		14.2 (1.5)	30.4 (8.0)
Casciato et al., 2024 [63]	52.5 (5.4) ^b	13.8 (1.2)	31.1 (7.9)
	0.0 ^c		
Chervonski et al., 2024 [25]		14.5 (2.2)	29.4 (10.8)
Cohen et al., 2024 [64]		13.3	35.3
Eng et al., 2025 [67]	43.4 ^b	13.4	
	0.0 ^c		
Fahy et al., 2024 [68]	55.4 ^b	14.7	32.5
Gajjar et al., 2024 [71]			36.3 (5.8)
Garg et al., 2024 [22]	4.4 (0.7) ^a	14.4 (1.5)	
Grippaudo et al., 2024 [72]	19.3 ^d		
Haidar et al., 2023 [24]	50.3 ^b	12.7 (1.9)	
Hurley et al., 2024 [74]	60.0 ^b		26.2
	0.0 ^c		
Ichhpujani et al., 2024 [75]		15.4 (1.8)	22.5 (7.0)
Johns et al., 2024 [77]	41.0 (4.3) ^b	12.0 (0.9)	

(Continued)

Table 3 (Continued)

LLM and educational materials	Quality	Readability	
	Mean (SD)	FKGL Mean (SD)	FRES Mean (SD)
Johns et al., 2024 [78]		11.5	
Lack et al., 2024 [82]	44.0 ^b	14.4	
Lauck et al., 2024 [85]		11.9 (1.6)	40.9 (6.9)
Lim et al., 2024 [86]	53.0 (2.1)	13.5 (1.4)	35.2 (7.4)
Lois et al., 2024 [87]		14.0	28.6
Sarikonda et al., 2024 [104]		15.0	27.5
Shaari et al., 2024 [105]		12.5	36.2
Snee et al., 2024 [107]		13.7 (1.2)	24.4 (10.6)
Wright et al., 2024 [111]		11.7	
Yan et al., 2024 [112]		13.6 (1.5)	
<i>ChatGPT-4.0</i>			
Fahy et al., 2024 [68]	62.1 ^b	13.7	28.1
Frodl et al., 2024 [70]	29.3 (5.8) ^d		
Gajjar et al., 2024 [71]			32.3 (5.8)
Lois et al., 2024 [87]	50.8 ^b	13.1	33.5
Mastrokostas et al., 2024 [88]	3.3 ^c	13.0 (2.3)	37.1 (10.3)
Momenaei et al., 2023 [91]		14.3	31.6
Polat et al., 2024 [97]		10.0 (1.9)	57.2 (10.7)
Robinson et al., 2024 [99]		10.3 (2.8)	
Sahin et al., 2024 [100]	45.4 (11.2) ^b	11.0 (2.3)	45.7 (9.7)
Sahin et al., 2024 [101]	50.7 (8.2) ^b	13.5 (1.6)	34.9 (7.3)
Shaari et al., 2024 [105]		11.7	37.9
Soroudi et al., 2024 [108]			40.4 (8.3)
Yan et al., 2024 [112]		14.2 (1.8)	
Yurdakurban et al., 2024 [113]	42.4 (7.6) ^d 3.8 (0.7) ^a		
<i>ChatGPT-3.0</i>			
Crook et al., 2023 [65]	58.0 ^a 0.0 ^c	14.5 (1.6)	33.5 (8.3)
Gajjar et al., 2024 [71]			30.7 (5.8)
Soroudi et al., 2024 [108]			31.5 (13.0)
<i>Google Gemini</i>			
Chervonski et al., 2024 [25]		8.2 (1.7)	58.9 (10.5)
Carlson et al., 2024 [62]		10.1 (1.9)	49.7 (9.6)
Garg et al., 2024 [22]	4.2 (0.9) ^a	9.9 (1.3)	
Ichhpujani et al., 2024 [75]		8.9 (1.2)	57.9 (7.1)
Lauck et al., 2024 [85]		9.0 (1.1)	60.6 (5.3)
Lim et al., 2024 [86]	49.1 (1.8) ^b	12.2 (2.1)	35.9 (9.8)
Lois et al., 2024 [87]		12.9	30.4
Robinson et al., 2024 [99]		9.3 (1.3)	
<i>Other LLMs</i>			
<i>Claude</i>			
Carlson et al., 2024 [62]		10.6 (1.5)	46.8 (8.2)
Lim et al., 2024 [86]	54.6 (2.2) ^b	11.2 (1.5)	45.3 (7.8)
Lois et al., 2024 [87]	48.0 ^b	11.0	39.1
<i>Bing Co-Pilot</i>			
Carlson et al., 2024 [62]		11.6 (1.9)	45.6 (10.3)
Lim et al., 2024 [86]	49.5 (10.7) ^b	10.6 (1.2)	39.9 (9.3)

(Continued)

Table 3 | (Continued)

LLM and educational materials	Quality	Readability	
	Mean (SD)	FKGL Mean (SD)	FRES Mean (SD)
<i>Perplexity</i>			
Carlson et al., 2024 [62]		13.3 (1.6)	36.4 (11.9)
<i>Medisearch</i>	4.1 (0.9) ^a		
Yurdakurban et al., 2024 [113]	47.1 (9.2) ^d		
<i>Openvidence</i>	4.1 (1.2) ^a		
Yurdakurban et al., 2024 [113]	48.8 (14.2) ^d		
<i>Google search</i>			
Alessandri-Bonetti et al., 2023 [20]	2.6 (1.3) ^a		
Cohen et al., 2024 [64]		6.7	59.8
Lauck et al., 2024 [85]		11.0 (3.5)	52.2 (14.9)
Mastrokostas et al., 2024 [88]	1.8 ^c	9.3 (2.4)	54.9 (14.2)
<i>Standard educational materials and provider responses</i>			
Casciato et al., 2024 [63]	66.9 (5.0) ^b	9.6 (1.2)	55.2 (7.3)
	2.0 (0.2) ^c		
Crook et al., 2023 [65]	71.0 ^b	8.6 (0.7)	60.4 (4.5)
	3.0 ^c		
Haidar et al., 2023 [24]	56.0 ^b	9.4 (1.9)	
Lois et al., 2024 [87]	63.5 ^b	11.2	43.6
Lois et al., 2024 [87]		9.2	55.2
Robinson et al., 2024 [99]		15.2 (5.9)	
Soroudi et al., 2024 [108]			53.2
Soroudi et al., 2024 [108]			53.5

DISCERN: Discerning the Quality of Health Information; EQIP: Ensuring Quality Information for Patients; FKGL: Flesch-Kincaid Grade Level; FRES: Flesch Reading Ease Score; GQS: Global Quality Score; JAMA: Journal of the American Medical Association Benchmark Criteria; SD: Standard Deviation. ^a: GQS ^b: DISCERN ^c: JAMA ^d: EQIP.

Table 4 | Comparison of accuracy, quality, and readability across ChatGPT versions, other LLMs, health LLMs and standard educational materials. This table summarizes results from weighted linear mixed-effects model with ChatGPT set as the reference group. Estimated marginal means were calculated for each group and Holm-adjusted post-hoc comparisons were performed to assess differences across model classifications

	Accuracy		Quality				Readability	
	Percentage accuracy score Mean (SE)	Percentage in top accuracy category (SE)	JAMA Mean (SE)	DISCERN Mean (SE)	GQS Mean (SE)	EQIP Mean (SE)	FKGL Mean (SE)	FRES Mean (SE)
	X ² (2) = 9.7; p = 0.008	X ² (1) = 5.2; p = 0.02	X ² (1) = 7.8; p = 0.005	X ² (2) = 24.0; p < 0.0001	X ² (2) = 57.7; p < 0.0001	X ² (1) = 6.3; p = 0.01	X ² (2) = 38.1; p < 0.0001	X ² (2) = 33.9; p < 0.0001
ChatGPT	74.7% (2.5) ^a	63.3% (5.3) ^b	0.3 (0.4) ^c	51.8 (1.8) ^d	4.3 (0.2) ^{f,g}	30.5 (6.7) ⁱ	13.4 (0.3) ^{j,k}	33.6 (1.5) ^{l,m}
Other LLMs [*]	65.9% (3.2) ^a	44.8% (7.9) ^b	–	49.7 (2.7) ^e	4.1 (0.2) ^{a,h}	–	10.5 (0.5) ^k	47.0 (2.5) ^{m,n}
Health LLMs	–	–	–	–	4.7 (0.2) ^{f,h}	36.1 (6.8) ⁱ	–	–
Standard educational materials	72.6 (5.6)	–	2.6 (0.6) ^c	65.6 (2.4) ^{d,e}	–	–	9.7 (0.6) ^j	54.8 (3.2) ^{l,n}

DISCERN: Discerning the Quality of Health Information; FKGL: Flesch-Kincaid Grade Level; FRES: Flesch Reading Ease Score; JAMA: Journal of the American Medical Association Benchmark Criteria; SE: Standard Error. ^{*}Other LLMs include Google Gemini, Bing Co-Pilot, Claude, CLOVA X, and Perplexity, grouped by metric: Percentage Accuracy Score (Google Gemini, Bing Co-Pilot, Claude, CLOVA X, Perplexity), Percentage in Top Accuracy Category (Google Gemini), DISCERN (Google Gemini, Claude and Bing-Co-Pilot), and FKGL/FRES (Google Gemini, Claude, Bing Co-Pilot, Perplexity). Dashes indicate that the model was not evaluated for that outcome. ^ap = 0.004; ^bp = 0.006; ^cp = 0.002; ^dp < 0.001; ^ep < 0.001; ^fp < 0.001; ^gp < 0.001; ^hp < 0.001; ⁱp = 0.0002; ^jp < 0.001; ^kp < 0.001; ^lp < 0.001; ^mp < 0.001; ⁿp = 0.04.

Table 5 Detailed comparison of accuracy, quality, and readability among ChatGPT versions, Google Gemini, other LLMs, Health LLMs and standard educational materials. This table summarizes results from weighted linear mixed-effects model with ChatGPT-3.5 set as the reference group. Estimated marginal means were calculated for each group and Holm-adjusted post-hoc comparisons were performed to assess differences across model classifications

	Accuracy		Quality				Readability	
	Percentage accuracy score Mean (SE)	Percentage in top accuracy category (SE)	JAMA Mean (SE)	DISCERN Mean (SE)	GQS Mean (SE)	EQIP Mean (SE)	FKGL Mean (SE)	FRES Mean (SE)
	$\chi^2(5) = 16.3$; $p = 0.006$	$\chi^2(2) = 9.0$; $p = 0.01$	$\chi^2(3) = 26.6$; $p < 0.0001$	$\chi^2(4) = 30.6$; $p < 0.0001$	$\chi^2(3) = 105.8$; $p < 0.0001$	$\chi^2(2) = 9.7$; $p = 0.008$	$\chi^2(5) = 46.9$; $p < 0.0001$	$\chi^2(7) = 44.0$; $p < 0.0001$
ChatGPT-3.0	67.7 (3.8)	–	–0.8 (0.2) ^{c, f, g}	–	–	–	15.3 (0.9) ^{v, x, y}	31.6 (4.1) ^{ee, ff}
ChatGPT-3.5	74.8 (2.8)	58.2 (5.9) ^a	0.2 (0.2) ^{c, d, e}	49.8 (2.0) ⁱ	4.5 (0.1) ^{m, n, o}	19.3 (9.5)	13.7 (0.4) ^{s, t, u}	31.2 (1.9) ^{bb, cc, dd}
ChatGPT-4.0	75.3 (3.0)	75.5 (8.1) ^b	3.3 (0.5) ^{d, f, h}	54.4 (2.2) ^j	3.8 (0.1) ^{m, p, q}	35.9 (6.7) ^r	12.5 (0.5) ^{z, aa}	38.0 (2.3) ^{gg, hh}
Google Gemini	66.3 (3.5)	40.8 (7.5) ^{a, b}	–	47.6 (5.3) ^k	4.3 (0.1) ^{n, p}	–	9.8 (0.5) ^{s, v, z}	49.0 (2.9) ^{bb, ee, gg}
Other LLMs [*]	65.7 (3.8)	–	–	50.8 (2.6) ^l	–	–	11.0 (0.7) ^{t, x}	43.1 (3.1) ^{cc}
Health LLMs	–	–	–	–	4.1 (0.1) ^{o, q}	41.6 (6.6) ^r	–	–
Standard educational materials	72.2 (5.4)	–	2.2 (0.2) ^{e, g, h}	66.0 (2.3) ^{i, j, k, l}	–	–	10.0 (0.6) ^{u, y, aa}	54.5 (3.0) ^{dd, ff, hh}

DISCERN: Discerning the Quality of Health Information; EQIP: Ensuring Quality Information for Patients; FKGL: Flesch-Kincaid Grade Level; FRES: Flesch Reading Ease Score; SE: Standard Error. *Other LLMs include Bing Co-Pilot, Claude, CLOVA X, and Perplexity grouped by metric: Percentage Accuracy Score (Bing Co-Pilot, Claude, CLOVA X, Perplexity), Percentage in Top Accuracy Category (Google Gemini), DISCERN (Claude, Bing-Co-Pilot), and FKGL/FRES (Claude, Bing Co-Pilot, Perplexity). Dashes indicate that the model was not evaluated for that outcome. ^a $p = 0.004$; ^b $p = 0.002$; ^c $p < 0.001$; ^d $p < 0.001$; ^e $p < 0.001$; ^f $p < 0.001$; ^g $p < 0.001$; ^h $p = 0.02$; ⁱ $p < 0.001$; ^j $p < 0.001$; ^k $p = 0.004$; ^l $p < 0.001$; ^m $p < 0.001$; ⁿ $p < 0.001$; ^o $p = 0.01$; ^p $p < 0.001$; ^q $p < 0.001$; ^r $p < 0.001$; ^s $p < 0.001$; ^t $p = 0.004$; ^u $p < 0.001$; ^v $p < 0.001$; ^w $p = 0.003$; ^x $p < 0.001$; ^y $p = 0.001$; ^z $p = 0.001$; ^{aa} $p = 0.01$; ^{bb} $p < 0.001$; ^{cc} $p = 0.006$; ^{dd} $p < 0.001$; ^{ee} $p = 0.006$; ^{ff} $p < 0.001$; ^{gg} $p = 0.02$; ^{hh} $p = 0.001$.

scientific references, Chelli et al., found that ChatGPT-3.5, GPT-4, and Google Bard exhibited hallucination rates of 39.6%, 28.6%, and 91.4%, respectively, indicating that LLM-produced references require thorough validation by researchers [125]. To address this issue, health-focused LLMs, such as MediSearch and OpenEvidence, have emerged as alternatives to circumvent this challenge and improve accuracy and reliability. Unlike general-purpose LLMs, which rely on broad internet-based datasets, these models prioritize high-quality, domain-specific knowledge to reduce misinformation and improve the accuracy of medical queries [113]. However, MediSearch has yet to gain widespread recognition, while OpenEvidence is only free and unlimited for verified U.S. healthcare providers, limiting broader public access [126,127]. As LLMs and generative AI technologies continue to evolve, it is crucial for both providers and patients to recognize that LLMs were originally created as a conversational tool designed to predict the most statistically likely sequence of words rather than verify factual accuracy [128]. However, future advancements in generative AI may improve source attribution, citation accuracy, and integration with validated medical databases, as seen with MediSearch and OpenEvidence, to improve the trustworthiness of LLM-generated content in surgical patient education.

Another consideration raised by our findings is that the reading level of LLM-generated responses are typically higher (often at a college-level reading level) than those of standard patient education materials. While the existing literature consistently reports this trend, the implications of these readability differences on patient comprehension, engagement, and health-decision making remain unclear. The American Medical

Association and National Institutes of Health recommend patient materials not exceed sixth- and eighth-grade reading levels, respectively [129]. However, many surgical subspecialty materials surpass these recommendations, often reaching a college-level reading grade [130,131]. Despite this, studies report that ChatGPT remains efficacious in answering surgical-related inquiries, raising questions about the extent to which readability influences patient uptake and usefulness [89,98]. Unlike static patient education materials, ChatGPT and other LLMs allow users to ask follow-up questions and refine prompts to receive more straightforward responses. This interactive and conversational component of LLMs may help compensate for the complex language by improving interface engagement and adaptability. Additionally, studies show that when prompted, ChatGPT can improve the readability of already existing surgical patient education materials, demonstrating its potential to tailor health information to diverse literacy needs [129,132]. Ultimately, further research is needed to assess patient interface and explore whether patients are able to understand accurate health information from LLM-generated responses.

While our study systematically evaluated three key performance metrics, additional factors warrant further investigation. Future research can address the reliability, consistency, and safety of LLM-generated information, and further clarify patient-centered outcomes such as satisfaction. Also, as LLMs are further employed in healthcare contexts, ensuring that their responses are not only accurate but also accessible to patients will be critical. Greater transparency may require compromise in the disclosure of back-end frameworks and structures of these LLMs [122].

This systematic review is subject to several limitations. First, while many included studies evaluated performance in answering questions, a significant portion of studies in our overall search did not explicitly or quantitatively evaluate key outcomes such as accuracy, quality, and readability thereby limiting the number of articles we could include in our analysis. Second, meaningful comparisons across questions subtypes were limited due to the variability in study and prompt design, as demonstrated by a mix of preoperative, postoperative, and procedural questions. Third, scoring reliability varied significantly across studies, and only some reported inter-rater reliability measures such as the intraclass correlation coefficients or kappa statistics. This raises concerns regarding the consistency and reproducibility of scoring methods used to evaluate LLM responses, which were sometimes assessed independently using custom Likert scales and at other times were evaluated via consensus-based assessments. Lastly, the wide range of employed outcome metrics complicated direct comparison across studies. Future research can prioritize standard outcome reporting and evaluation criteria, along with the development of appropriate risk of bias assessment tools, to strengthen comparability and reliability across studies regarding the effectiveness of LLMs in surgical patient education.

5. CONCLUSION

The systematic review performed within this study demonstrated that LLM-generated responses were generally accurate compared to standard patient education materials. While ChatGPT outperformed other LLMs in accuracy and operated at a college-level reading comprehension, it did not consistently match the quality of standard patient education materials. Further refinements in AI-generated health content, including improvements in clarity, reliability, surgery-specific accuracy, will be essential to strengthen its effectiveness as a patient education tool.

DECLARATIONS

Conflict of Interest

The authors declare that they have no conflicts of interest.

Authors' Contribution

AEH and JCL contributed in concept and design. AEH, AKP, AAA, and ATP contributed in acquisition of data. AEH and AKP contributed in analysis of data. AEH, AKP, and KS contributed in interpretation of data and also in drafting the manuscript. JCL contributed in funding and supervision of the project. AEH, AKP, and JCL contributed in administrative, technical or material support of the project. All authors are involved in the critical revision of the manuscript for important intellectual content.

Funding/Support

This work was supported by the Bernard G. Sarnat Endowment for Craniofacial Biology (JCL), the Jean Perkins Foundation (JCL), and the National Center for Advancing Translational Science (NCATS) of the National Institutes of Health under the

UCLA Clinical and Translational Science Institute grant number UL1TR001881.

Role of Funder/Sponsor

The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Supplementary Materials

Supplementary data related to this article can be found at <https://doi.org/10.55578/joaims.251202.001>.

REFERENCES

- [1] Y.K. Dwivedi, L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, et al., Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy, *Int. J. Inf. Manag.* 57 (2021), 101994.
- [2] A. Bordas, P. Le Masson, M. Thomas, B. Weil, What is generative in generative artificial intelligence? A design-based perspective, *Res. Eng. Des.* 35 (2024), 427–443.
- [3] J.A. Jindal, M.P. Lungren, N.H. Shah, Ensuring useful adoption of generative artificial intelligence in healthcare, *J. Am. Med. Inform. Assoc.* 31 (2024), 1441–1444.
- [4] P. Yu, H. Xu, X. Hu, C. Deng, Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration, *Healthcare*. 11 (2023), 2776.
- [5] S.P. TerKonda, A.A. TerKonda, J.M. Sacks, B.M. Kinney, G.C. Gurtner, J.M. Nachbar, et al., Artificial intelligence: singularity approaches, *Plast. Reconstr. Surg.* 153 (2024), 204e–217e.
- [6] S. Qin, B. Chislett, J. Ischia, W. Ranasinghe, D. de Silva, J. Coles-Black, et al., ChatGPT and generative AI in urology and surgery—a narrative review, *BJUI Compass*. 5 (2024), 927–935.
- [7] S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, et al., Testing and evaluation of health care applications of large language models: a systematic review, *JAMA*. 333 (2025), 319–328.
- [8] J. He, S.L. Baxter, J. Xu, J. Xu, X. Zhou, K. Zhang, The practical implementation of artificial intelligence technologies in medicine, *Nat. Med.* 25 (2019), 30–36.
- [9] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, *Cell*. 172 (2018), 1122–1131.e9.
- [10] T. Nadarzynski, O. Miles, A. Cowie, D. Ridge, Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study, *Digit. Health*. 5 (2019), 2055207619871808.
- [11] A.A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B.M. Bewick, M. Househ, Perceptions and opinions of patients about mental health chatbots: scoping review, *J. Med. Internet Res.* 23 (2021), e17828.
- [12] K. Kretzschmar, H. Tyroll, G. Pavarini, A. Manzini, I. Singh, NeurOx Young People's Advisory Group, Can your phone be your therapist? Young people's ethical perspectives on the

- use of fully automated conversational agents (Chatbots) in mental health support, *Biomed. Inform. Insights*. 11 (2019), 1178222619829083.
- [13] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir, et al., Conversational agents in healthcare: a systematic review, *J. Am. Med. Inform. Assoc.* 25 (2018), 1248–1258.
 - [14] S.-J. Lin, C.-Y. Sun, D.-N. Chen, Y.-N. Kang, N.M. Lai, K.-H. Chen, et al., Perioperative application of chatbots: a systematic review and meta-analysis, *BMJ Health Care Inform.* 31 (2024), e100985.
 - [15] M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, B. Lebouché, Roles, users, benefits, and limitations of chatbots in health care: rapid review, *J. Med. Internet Res.* 26 (2024), e56930.
 - [16] ChatGPT | OpenAI, ChatGPT overview, <https://openai.com/chatgpt/overview/>.
 - [17] M. Javaid, A. Haleem, R.P. Singh, ChatGPT for healthcare services: an emerging stage for an innovative perspective, *BenchCouncil Trans. Benchmarks Stand. Eval.* 3 (2023), 100105.
 - [18] Y. Mu, D. He, The potential applications and challenges of ChatGPT in the medical field, *Int. J. Gen. Med.* 17 (2024), 817–826.
 - [19] M. Yahagi, R. Hiruta, C. Miyauchi, R.N. Shoko Tanaka, R.N. Aya Taguchi, Y. Yaguchi, Comparison of conventional anesthesia nurse education and an artificial intelligence chatbot (ChatGPT) intervention on preoperative anxiety: a randomized controlled trial, *J. Perianesth. Nurs.* 39 (2024), 767–771.
 - [20] M. Alessandri-Bonetti, H.Y. Liu, M. Palmesano, V.T. Nguyen, F.M. Egro, Online patient education in body contouring: a comparison between Google and ChatGPT, *J. Plast. Reconstr. Aesthet. Surg.* 87 (2023), 390–402.
 - [21] Y. Balel, Can ChatGPT be used in oral and maxillofacial surgery?, *J. Stomatol. Oral Maxillofac. Surg.* 124 (2023), 101471.
 - [22] N. Garg, D.J. Campbell, A. Yang, A. McCann, A.E. Moroco, L.E. Estephan, et al., Chatbots as patient education resources for aesthetic facial plastic surgery: evaluation of ChatGPT and Google Bard responses, *Facial Plast. Surg. Aesthet. Med.* 26 (2024), 665–673.
 - [23] A.T. Anastasio, F.B. Mills, M.P. Karavan, S.B. Adams, Evaluating the quality and usability of artificial intelligence-generated responses to common patient questions in foot and ankle surgery, *Foot Ankle Orthop.* 8 (2023), 24730114231209919.
 - [24] O. Haidar, A. Jaques, P.W. McCaughran, M.J. Metcalfe, AI-generated information for vascular patients: assessing the standard of procedure-specific information provided by the ChatGPT AI-language model, *Cureus*. 15 (2023), e49764.
 - [25] E. Chervonski, K.B. Harish, C.B. Rockman, M. Sadek, K.A. Teter, G.R. Jacobowitz, et al., Generative artificial intelligence chatbots may provide appropriate informational responses to common vascular surgery questions by patients, *Vascular*. 33 (2024), 229–237.
 - [26] Y.-X. Sun, Z.-M. Li, J.-Z. Huang, N.-z. Yu, X. Long, GPT-4: the future of cosmetic procedure consultation?, *Aesthet. Surg. J.* 43 (2023), NP670–NP672.
 - [27] Y. Xie, I. Seth, D.J. Hunter-Smith, W.M. Rozen, R. Ross, M. Lee, Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT, *Aesth. Plast. Surg.* 47 (2023), 1985–1993.
 - [28] I. Patel, A. Om, D. Cuzzzone, G. Garcia Nores, Comparing ChatGPT vs surgeon-generated informed consent documentation for plastic surgery procedures, *Aesth. Surg. J. Open Forum*. 6 (2024), ojae092.
 - [29] L.A. Vaira, J.R. Lechien, A. Maniaci, G. Tanda, V. Abbate, F. Allevi, et al., Evaluating AI-generated informed consent documents in oral surgery: a comparative study of ChatGPT-4, Bard gemini advanced, and human-written consents, *J. Cranio-Maxillofac. Surg.* 53 (2025), 18–23.
 - [30] H. Decker, K. Trang, J. Ramirez, A. Colley, L. Pierce, M. Coleman, et al., Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures, *JAMA Netw. Open*. 6 (2023), e2336997.
 - [31] J. Abi-Rafteh, B. Bassiri-Tehrani, R. Kazan, S.A. Hanna, J. Kanevsky, F. Nahai, Comparative performance of current patient-accessible artificial intelligence large language models in the preoperative education of patients in facial aesthetic surgery, *Aesth. Surg. J. Open Forum*. 6 (2024), ojae058.
 - [32] T.B. Amen, K.A. Torabian, T. Subramanian, B.W. Yang, A. Liimakka, D. Fufa, Quality of ChatGPT responses to frequently asked questions in carpal tunnel release surgery, *Plast. Reconstr. Surg. Glob. Open*. 12 (2024), e5822.
 - [33] T.S. Kim, C.T. Yu, C. Hinson, E. Fung, O. Allam, R.S. Nazerali, et al., ChatGPT virtual assistant for breast reconstruction: assessing preferences for a traditional chatbot versus a human AI VideoBot, *Plast. Reconstr. Surg. Glob. Open*. 12 (2024), e6202.
 - [34] W. Li, Y. Zhang, F. Chen, ChatGPT in colorectal surgery: a promising tool or a passing fad?, *Ann. Biomed. Eng.* 51 (2023), 1892–1897.
 - [35] F.K. Aljindan, M.H. Shawosh, L. Altamimi, S. Arif, H. Mortada, Utilization of ChatGPT-4 in plastic and reconstructive surgery: a narrative review, *Plast. Reconstr. Surg. Glob. Open*. 11 (2023), e5305.
 - [36] J. Abi-Rafteh, N. Henry, H.H. Xu, B. Bassiri-Tehrani, A. Arezki, R. Kazan, et al., Utility and comparative performance of current artificial intelligence large language models as post-operative medical support chatbots in aesthetic surgery, *Aesthet. Surg. J.* 44 (2024), 889–896.
 - [37] M. Capelleras, G.A. Soto-Galindo, M. Cruellas, F. Apaydin, ChatGPT and rhinoplasty recovery: an exploration of AI's role in postoperative guidance, *Facial Plast. Surg.* 40 (2024), 628–631.
 - [38] S.A. Albehairi, O.M. Alsahli, L.F. Bander, T.M. Albilasi, E.S. Aljardani, Postoperative otoplasty care with ChatGPT-4: a study on artificial intelligence (AI)-assisted patient concern and education, *J. Craniofac. Surg.* 36 (2025), 296–298.
 - [39] E. Kaba, M. Beyazal, F.B. Çeliker, İ. Yel, T.J. Vogl, Accuracy and readability of ChatGPT on potential complications of interventional radiology procedures: AI-powered patient interviewing, *Acad. Radiol.* 32 (2025), 1547–1553.
 - [40] S.M. Staubli, H.L. Walker, F. Saner, C.H. Salinas, D.C. Broering, M. Malagò, et al., Decoding the Clavien-Dindo classification: artificial intelligence (AI) as a novel tool to grade post-operative complications, *BJS*. 111 (2024), znae118.031.
 - [41] G.A. Soto-Galindo, M. Capelleras, M. Cruellas, F. Apaydin, Effectiveness of ChatGPT in identifying and accurately guiding patients in rhinoplasty complications, *Facial Plast. Surg.* 40 (2024), 623–627.
 - [42] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, *Nat. Med.* 29 (2023), 1930–1940.
 - [43] D. Busch, C. Za'in, H.M. Chan, A. Haryanto, W. Agustiono, K. Yu, et al., A blueprint for large language model-augmented telehealth for HIV mitigation in Indonesia: a scoping review of a novel therapeutic modality, *Health Inform. J.* 31 (2025), 14604582251315595.

- [44] E. Roldan-Vasquez, S. Mitri, S. Bhasin, T. Bharani, K. Capasso, M. Haslinger, et al., Reliability of artificial intelligence chatbot responses to frequently asked questions in breast surgical oncology, *J. Surg. Oncol.* 130 (2024), 188–203.
- [45] V.K. Viswanathan, V.K. Jain, A. Vaish, M. Jeyaraman, K.P. Iyengar, R. Vaishya, Chatbots and their applications in medical fields: current status and future trends: a scoping review, *Apollo Med.* 21 (2024), 386–392.
- [46] P. Lee, S. Bubeck, J. Petro, Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine, *N. Engl. J. Med.* 388 (2023), 1233–1239.
- [47] X.T. Gonzalez, K. Steger-May, J. Abraham, Just another tool in their repertoire: uncovering insights into public and patient perspectives on clinicians' use of machine learning in perioperative care, *J. Am. Med. Inform. Assoc.* 32 (2025), 150–162.
- [48] A. Palanica, P. Flaschner, A. Thommandram, M. Li, Y. Fossat, Physicians' perceptions of chatbots in health care: cross-sectional web-based survey, *J. Med. Internet Res.* 21 (2019), e12887.
- [49] A.J. Thirunavukarasu, R. Hassan, S. Mahmood, R. Sanghera, K. Barzangi, M. El Mukashfi, et al., Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care, *JMIR Med. Educ.* 9 (2023), e46599.
- [50] A. Gilson, C.W. Safranek, T. Huang, V. Socrates, L. Chi, R.A. Taylor, et al., How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment, *JMIR Med. Educ.* 9 (2023), e45312.
- [51] S.S. Hubany, F.D. Scala, K. Hashemi, S. Kapoor, J.R. Fedorova, M.J. Vaccaro, et al., ChatGPT-4 surpasses residents: a study of artificial intelligence competency in plastic surgery in-service examinations and its advancements from ChatGPT-3.5, *Plast. Reconstr. Surg. Glob. Open.* 12 (2024), e6136.
- [52] D. Charnock, S. Shepperd, G. Needham, R. Gann, DISCERN: an instrument for judging the quality of written consumer health information on treatment choices, *J. Epidemiol. Community Health.* 53 (1999), 105–111.
- [53] W.M. Silberg, G.D. Lundberg, R.A. Musacchio, Assessing, controlling, and assuring the quality of medical information on the internet: Caveant lector et viewer—let the reader and viewer beware, *JAMA.* 277 (1997), 1244–1245.
- [54] B. Moul, L.S. Franck, H. Brady, Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information, *Health Expect.* 7 (2004), 165–175.
- [55] A. Bernard, M. Langille, S. Hughes, C. Rose, D. Leddin, S. Veldhuyzen van Zanten, A systematic review of patient inflammatory bowel disease information resources on the World Wide Web, *Am. J. Gastroenterol.* 102 (2007), 2070–2077.
- [56] J.P. Kincaid, R.P. Fishburne, R.L. Rogers, B.S. Chissom, Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Defense Technical Information Center, Chief of Naval Technical Training, Naval Air Station Memphis, Millington, Springfield, Distributed by NTIS, 1975.
- [57] R. Flesch, A new readability yardstick, *J. Appl. Psychol.* 32 (1948), 221–233.
- [58] J.M. Adelstein, M.A. Sinkler, L.T. Li, R. Chen, R.J. Gillespie, J. Calcei, Assessing ChatGPT responses to frequently asked questions regarding total shoulder arthroplasty, *Semin. Arthroplasty.* 34 (2024), 416–424.
- [59] E.M. Al-Sharif, R.C. Penteado, N. Dib El Jalbout, N.J. Topilow, M.K. Shoji, D.O. Kikkawa, et al., Evaluating the accuracy of ChatGPT and Google BARD in fielding oculoplastic patient queries: a comparative study on artificial versus human intelligence, *Ophthalmic. Plast. Reconstr. Surg.* 40 (2024), 303–311.
- [60] Y. AlShehri, M. McConkey, P. Lodhia, ChatGPT provides satisfactory but occasionally inaccurate answers to common patient hip arthroscopy questions, *Arthrosc.* 41 (2025), 1337–1347.
- [61] S.S. Bains, J.A. Dubin, D. Hameed, O.C. Sax, S. Douglas, M.A. Mont, et al., Use and application of large language models for patient questions following total knee arthroplasty, *J. Arthroplasty.* 39 (2024), 2289–2294.
- [62] J.A. Carlson, R.Z. Cheng, A. Lange, N. Nagalakshmi, J. Rabets, T. Shah, et al., Accuracy and readability of artificial intelligence chatbot responses to vasectomy-related questions: public beware, *Cureus.* 16 (2024), e67996.
- [63] D. Casciato, S. Mateen, S. Cooperman, D. Pesavento, R.A. Brandao, Evaluation of online AI-generated foot and ankle surgery information, *J. Foot Ankle Surg.* 63 (2024), 680–683.
- [64] S.A. Cohen, A. Brant, A.C. Fisher, S. Pershing, D. Do, C. Pan, Dr. Google vs. Dr. ChatGPT: exploring the use of artificial intelligence in ophthalmology by comparing the accuracy, safety, and readability of responses to frequently asked patient questions regarding cataracts and cataract surgery, *Semin. Ophthalmol.* 39 (2024), 472–479.
- [65] B.S. Crook, C.N. Park, E.T. Hurley, M.J. Richard, T.S. Pidgeon, Evaluation of online artificial intelligence-generated information on common hand procedures, *J. Hand Surg. Am.* 48 (2023), 1122–1127.
- [66] Y. Endo, K. Sasaki, Z. Moazzam, H.A. Lima, A. Schenk, A. Limkemann, et al., Quality of ChatGPT responses to questions related to liver transplantation, *J. Gastrointest. Surg.* 27 (2023), 1716–1719.
- [67] E. Eng, C. Mowers, D. Sachdev, P. Yerke-Hansen, G.R. Jackson, D.M. Knapik, et al., Chat Generative Pre-Trained Transformer (ChatGPT) – 3.5 responses require advanced readability for the general population and may not effectively supplement patient-related information provided by the treating surgeon regarding common questions about rotator cuff repair, *Arthrosc.* 41 (2025), 42–52.
- [68] S. Fahy, S. Oehme, D. Milinkovic, T. Jung, B. Bartek, Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament injury, *J. Pers. Med.* 14 (2024), 104.
- [69] D. Ferrari-Light, R.E. Merritt, D. D'Souza, M.K. Ferguson, S. Harrison, M.L. Madariaga, et al., Evaluating ChatGPT as a patient resource for frequently asked questions about lung cancer surgery—a pilot study, *J. Thorac. Cardiovasc. Surg.* 169 (2025), 1174–1180.e18.
- [70] A. Frodl, A. Fuchs, T. Yilmaz, K. Izadpanah, H. Schmal, M. Siegel, ChatGPT as a source for patient information on patellofemoral surgery—a comparative study among laymen, doctors, and experts, *Clin. Pract.* 14 (2024), 2376–2384.
- [71] A.A. Gajjar, R.P. Kumar, E.D. Paliwoda, C.C. Kuo, S. Adida, A.D. Legarreta, et al., Usefulness and accuracy of artificial intelligence Chatbot responses to patient questions for neurosurgical procedures, *Neurosurgery.* 95 (2024) 171–178.

- [72] F.R. Grippaudo, S. Nigrelli, A. Patrignani, D. Ribuffo, Quality of the information provided by ChatGPT for patients in breast plastic surgery: are we already in the future?, *JPRAS Open*. 40 (2024), 99–105.
- [73] X. Hu, M. Niemann, A. Kienzle, K. Braun, D.A. Back, C. Gwinner, et al., Evaluating ChatGPT responses to frequently asked patient questions regarding periprosthetic joint infection after total hip and knee arthroplasty, *Digit. Health*. 10 (2024), 20552076241272620.
- [74] E.T. Hurley, B.S. Crook, S.G. Lorentz, R.M. Danilkowicz, B.C. Lau, D.C. Taylor, et al., Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery, *Arthrosc*. 40 (2024), 726–731.e6.
- [75] P. Ichhpujani, U.P.S. Parmar, S. Kumar, Appropriateness and readability of Google Bard and ChatGPT-3.5 generated responses for surgical treatment of glaucoma, *Rom. J. Ophthalmol*. 68 (2024), 243–248.
- [76] M.H. Jo, M.-J. Kim, H.-K. Oh, M.J. Choi, H.-R. Shin, T.-G. Lee, et al., Communicative competence of generative artificial intelligence in responding to patient queries about colorectal cancer surgery, *Int. J. Colorectal Dis*. 39 (2024), 94–101.
- [77] W.L. Johns, B.J. Martinazzi, B. Miltenberg, H.H. Nam, S. Hammoud, ChatGPT provides unsatisfactory responses to frequently asked questions regarding anterior cruciate ligament reconstruction, *Arthrosc*. 40 (2024), 2067–2079.e1.
- [78] W.L. Johns, A. Kellish, D. Farronato, M.G. Ciccotti, S. Hammoud, ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction, *Arthrosc. Sports Med. Rehabil*. 6 (2024), 100893.
- [79] J. Kaarre, R. Feldt, L.E. Keeling, S. Dadoo, B. Zsidai, J.D. Hughes, et al., Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information, *Knee Surg. Sports Traumatol. Arthrosc*. 31 (2023), 5190–5198.
- [80] J. Kaarre, R. Feldt, B. Zsidai, E.H. Senorski, E.M. Rydberg, O. Wolf, et al., ChatGPT can yield valuable responses in the context of orthopaedic trauma surgery, *J. Exp. Orthop*. 11 (2024), e12047.
- [81] A. Kasapovic, T. Ali, M. Babasiz, J. Bojko, M. Gathen, R. Kaczmarczyk, et al., Does the information quality of ChatGPT meet the requirements of orthopedics and trauma surgery?, *Cureus*. 16 (2024), e60318.
- [82] B.T. Lack, E. Mouhawasse, J.T. Childers, G.R. Jackson, S.V. Daji, P. Yerke-Hansen, et al., Can ChatGPT answer patient questions regarding reverse shoulder arthroplasty?, *J. ISAKOS*. 9 (2024), 100323.
- [83] S.P. Lang, E.T. Yoseph, A.D. Gonzalez-Suarez, R. Kim, P. Fatemi, K. Wagner, et al., Analyzing large language models' responses to common lumbar spine fusion surgery questions: a comparison between ChatGPT and Bard, *Neurospine*. 21 (2024), 633–641.
- [84] S. Lang, J. Vitale, T.F. Fekete, D. Haschtmann, R. Reitmeir, M. Ropelato, et al., Are large language models valid tools for patient information on lumbar disc herniation? The spine surgeons' perspective, *Brain Spine*. 4 (2024), 102804.
- [85] K.C. Lauck, S.W. Cho, M. DaCunha, J. Wuennenberg, S. Aasi, M. Alam, et al., The utility of artificial intelligence platforms for patient-generated questions in Mohs micrographic surgery: a multi-national, blinded expert panel evaluation, *Int. J. Dermatol*. 63 (2024), 1592–1598.
- [86] B. Lim, I. Seth, R. Cuomo, P.S. Kenney, R.J. Ross, F. Sofiadellis, et al., Can AI answer my questions? Utilizing artificial intelligence in the perioperative assessment for abdominoplasty patients, *Aesthetic Plast. Surg*. 48 (2024), 4712–4724.
- [87] A. Lois, R. Yates, M. Ivy, C. Inaba, R. Tatum, L. Cetrulo, et al., Accuracy of natural language processors for patients seeking inguinal hernia information, *Surg. Endosc*. 38 (2024), 7409–7415.
- [88] P.G. Mastrokostas, L.E. Mastrokostas, A.K. Emara, I.J. Wellington, E. Ginalis, J.K. Houten, et al., GPT-4 as a source of patient information for anterior cervical discectomy and fusion: a comparative analysis against Google Web Search, *Glob. Spine J*. 14 (2024), 2389–2398.
- [89] A.P. Mika, J.R. Martin, S.M. Engstrom, G.G. Polkowski, J.M. Wilson, Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty, *J. Bone Joint Surg. Am*. 105 (2023), 1519–1526.
- [90] A.P. Mika, H.E. Mulvey, S.M. Engstrom, G.G. Polkowski, J.R. Martin, J.M. Wilson, Can ChatGPT answer patient questions regarding total knee arthroplasty?, *J. Knee Surg*. 37 (2024), 664–673.
- [91] B. Momenaei, T. Wakabayashi, A. Shahlaee, A.F. Durrani, S.A. Pandit, K. Wang, et al., Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases, *Ophthalmol. Retina*. 7 (2023), 862–868.
- [92] E. Mouhawasse, C.W. Haff, P. Kumar, B. Lack, K. Chu, U. Bansal, Can AI chatbots accurately answer patient questions regarding vasectomies?, *Int. J. Impot. Res*. (2024), Online ahead of print.
- [93] M.M. Munir, Y. Endo, A. Ejaz, M. Dillhoff, J.M. Cloyd, T.M. Pawlik, Online artificial intelligence platforms and their applicability to gastrointestinal surgical operations, *J. Gastrointest. Surg*. 28 (2024), 64–69.
- [94] P. Nwokeji, P. Cannata, A. Tadinada, C. Lottinger, A. Perez-Pino, The effectiveness of artificial intelligence-based virtual assistants in addressing frequently encountered inquiries from patients in oral surgery clinics, *J. Oral Maxillofac. Surg*. 82 (2024), S24–S25.
- [95] S.F. Oliveria, R.L. Rodriguez, D. Bowers, D. Kantor, J.D. Hilliard, E.H. Monari, et al., Safety and efficacy of dual-lead thalamic deep brain stimulation for patients with treatment-refractory multiple sclerosis tremor: a single-centre, randomised, single-blind pilot trial, *Lancet Neurol*. 16 (2017), 691–700.
- [96] N.S. Patil, R. Huang, A. Mihalache, E. Kisilevsky, J. Kwok, M.M. Popovic, et al., The ability of artificial intelligence chatbots ChatGPT and Google Bard to accurately convey preoperative information for patients undergoing ophthalmic surgeries, *Retina*. 44 (2024), 950–953.
- [97] E. Polat, Y.B. Polat, E. Senturk, R. Dogan, A. Yenigun, S. Tugrul, et al., Evaluating the accuracy and readability of ChatGPT in providing parental guidance for adenoidectomy, tonsillectomy, and ventilation tube insertion surgery, *Int. J. Pediatr. Otorhinolaryngol*. 181 (2024), 111998.
- [98] S. Reddy, B. Harnisch, S. Honig, K. Rotker, High capability of ChatGPT to answer common patient questions on vasectomy reviewed by sexual medicine specialists, *J. Sex Med*. 21 (2024), qdae001.361.

- [99] M.A. Robinson, M. Belzberg, S. Thakker, K. Bibee, E. Merkel, D.F. MacFarlane, et al., Assessing the accuracy, usefulness, and readability of artificial-intelligence-generated responses to common dermatologic surgery questions for patient education: a double-blinded comparative study of ChatGPT and Google Bard, *J. Am. Acad. Dermatol.* 90 (2024), 1078–1080.
- [100] Ş. Şahin, M.S. Tekin, Y.E. Yigit, B. Erkmen, Y.K. Duymaz, İ. Bahşi, Evaluating the success of ChatGPT in addressing patient questions concerning thyroid surgery, *J. Craniofac. Surg.* 35 (2024), e572–e575.
- [101] S. Sahin, B. Erkmen, Y.K. Duymaz, F. Bayram, A.M. Tekin, V. Topsakal, Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery, *Front. Surg.* 11 (2024), 1373843.
- [102] J.S. Samaan, Y.H. Yeo, N. Rajeev, L. Hawley, S. Abel, W.H. Ng, et al., Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery, *Obes. Surg.* 33 (2023), 1790–1796.
- [103] A.P. Samsonov, A. Habibi, J.J. Butler, R.J. Walls, J.G. Kennedy, Artificial intelligence language models are useful tools for patients undergoing total ankle replacement, *Foot Ankle Spec.* (2024), 19386400241249810.
- [104] A. Sarikonda, R. Abishek, E.L. Isch, A.A. Momin, M. Self, A. Sambangi, et al., Assessing the clinical appropriateness and practical utility of ChatGPT as an educational resource for patients considering minimally invasive spine surgery, *Cureus.* 16 (2024), e71105.
- [105] A.L. Shaari, A.N. Fano, O. Anakwenze, C. Klifto, Appraisal of ChatGPT's responses to common patient questions regarding Tommy John surgery, *Shoulder Elb.* 16 (2024), 429–435.
- [106] A.M. Smith, E.A. Jacquez, E.H. Argintar, Assessing the efficacy of an AI-powered chatbot (ChatGPT) in providing information on orthopedic surgeries: a comparative study with expert opinion, *Cureus.* 16 (2024), e63287.
- [107] I. Snee, C.X. Lava, K.R. Li, G.D. Corral, The utility of ChatGPT in gender-affirming mastectomy education, *J. Plast. Reconstr. Aesthet. Surg.* 99 (2024), 432–435.
- [108] D. Soroudi, A. Gozali, J.A. Knox, N. Parmeshwar, R. Sadjadi, J.C. Wilson, et al., Comparing provider and ChatGPT responses to breast reconstruction patient questions in the electronic health record, *Ann. Plast. Surg.* 93 (2024), 541–545.
- [109] J.B. Villarreal-Espinosa, R.S. Berreta, F. Allende, J.R. Garcia, S. Ayala, F. Familiari, et al., Accuracy assessment of ChatGPT responses to frequently asked questions regarding anterior cruciate ligament surgery, *Knee.* 51 (2024), 84–92.
- [110] C.A. White, Y.A. Masturov, E. Haunschild, E. Michaelson, D.R. Shukla, P.J. Cagle, Can ChatGPT reliably answer the most common patient questions regarding total shoulder arthroplasty?, *J. Shoulder Elbow Surg.* 34 (2025), e254–e264.
- [111] B.M. Wright, M.S. Bodnar, A.D. Moore, M.C. Maseda, M.P. Kucharik, C.C. Diaz, et al., Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients?, *Bone Jt. Open.* 5 (2024), 139–146.
- [112] S.-Y. Yan, Y.-F. Liu, L. Ma, L.-L. Xiao, X. Hu, R. Guo, et al., Walking forward or on hold: could the ChatGPT be applied for seeking health information in neurosurgical settings?, *Ibrain.* 10 (2024), 111–115.
- [113] E. Yurdakurban, K.G. Topsakal, G.S. Duran, A comparative analysis of AI-based chatbots: assessing data quality in orthognathic surgery related patient information, *J. Stomatol. Oral Maxillofac. Surg.* 125 (2024), 101757.
- [114] S. Zhang, Z.Q.G. Liao, K.L.M. Tan, W.L. Chua, Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement, *Knee Surg. Relat. Res.* 36 (2024), 15.
- [115] A.Z. Fazilat, C.E. Berry, A. Churukian, C. Lavin, L. Kamen, C. Brenac, et al., AI-based cleft lip and palate surgical information is preferred by both plastic surgeons and patients in a blind comparison, *Cleft Palate Craniofac. J.* 62 (2025), 1542–1548.
- [116] C.E. Berry, A.Z. Fazilat, C. Lavin, H. Lintel, N. Cole, C.S. Stingl, et al., Both patients and plastic surgeons prefer artificial intelligence-generated microsurgical information, *J. Reconstr. Microsurg.* 40 (2024), 657–664.
- [117] J.R. Lechien, A. Rameau, Applications of ChatGPT in otolaryngology–head neck surgery: a state of the art review, *Otolaryngol. Head Neck Surg.* 171 (2024), 667–677.
- [118] Q. Wei, Z. Yao, Y. Cui, B. Wei, Z. Jin, X. Xu, Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis, *J. Biomed. Inform.* 151 (2024), 104620.
- [119] D. Chen, R.S. Huang, J. Jomy, P. Wong, M. Yan, J. Croke, et al., Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases, *JAMA Netw. Open.* 7 (2024), e2437711.
- [120] A. Roman, L. Al-Sharif, M. Al Gharyani, The expanding role of ChatGPT (Chat-Generative Pre-Trained Transformer) in neurosurgery: a systematic review of literature and conceptual framework, *Cureus.* 15 (2023), e43502.
- [121] S.M. Pressman, S. Borna, C.A. Gomez-Cabello, S.A. Haider, C.R. Haider, A.J. Forte, Clinical and surgical applications of large language models: a systematic review, *J. Clin. Med.* 13 (2024), 3041.
- [122] B. Huo, A. Boyle, N. Marfo, W. Tangamornsuksan, J.P. Steen, T. McKechnie, et al., Large language models for chatbot health advice studies: a systematic review, *JAMA Netw. Open.* 8 (2025), e2457879.
- [123] L. Wang, Z. Wan, C. Ni, Q. Song, Y. Li, E. Clayton, et al., Applications and concerns of ChatGPT and other conversational large language models in healthcare: systematic review, *J. Med. Internet Res.* 26 (2024), e22769.
- [124] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, et al., A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions, *ACM Trans. Inf. Syst.* 43 (2025), 1–55.
- [125] M. Chelli, J. Descamps, V. Lavoué, C. Trojani, M. Azar, M. Deckert, et al., Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis, *J. Med. Internet Res.* 26 (2024), e53164.
- [126] MediSearch, MediSearch platform, <https://medisearch.io>.
- [127] OpenEvidence, OpenEvidence platform, <https://www.openevidence.com>.
- [128] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, et al., Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond, *ACM Trans. Knowl. Discov. Data.* 18 (2024), 1–32.
- [129] M. Vallurupalli, N.D. Shah, R.M. Vyas, Validation of ChatGPT 3.5 as a tool to optimize readability of patient-facing craniofacial education materials, *Plast. Reconstr. Surg. Glob. Open.* 12 (2024), e5575.

- [130] A. Phan, A. Jubril, E. Menga, A. Mesfin, Readability of the most commonly accessed online patient education materials pertaining to surgical treatments of the spine, *World Neurosurg.* 152 (2021), e583–e588.
- [131] K.C. Lee, E.T. Berg, H.E. Jazayeri, S.-K. Chuang, S.B. Eisig, Online patient education materials for orthognathic surgery fail to meet readability and quality standards, *J. Oral Maxillofac. Surg.* 77 (2019), 180.e1–180.e8.
- [132] G.J. Kirchner, R.Y. Kim, J.B. Weddle, J.E. Bible, Can artificial intelligence improve the readability of patient education materials?, *Clin. Orthop. Relat. Res.* 481 (2023), 2260–2267.