BMC Medical Education

RESEARCH

Open Access

Large language models versus traditional textbooks: optimizing learning for plastic surgery case preparation

Chandler Hinson[1,3*], Cybil Sierra Stingl[2] and Rahim Nazerali[2]

Abstract

Background Large language models (LLMs), such as ChatGPT-4 and Gemini, represent a new frontier in surgical education by offering dynamic, interactive learning experiences. Despite their potential, concerns about the accuracy, depth of knowledge, and bias in LLM responses persist. This study evaluates the effectiveness of LLMs in aiding surgical trainees in plastic and reconstructive surgery through comparison with traditional case-preparation textbooks.

Methods Six representative cases from key areas of plastic and reconstructive surgery—craniofacial, hand, microsurgery, burn, gender-affirming, and aesthetics—were selected. Four types of questions were developed for each case to cover clinical anatomy, indications, contraindications, and complications. Responses from LLMs (ChatGPT-4 and Gemini) and textbooks were compared using surveys distributed to medical students, research fellows, residents, and attending surgeons. Reviewers rated each response on accuracy, thoroughness, usefulness for case preparation, brevity, and overall quality using a 5-point Likert scale. Statistical analyses, including ANOVA and unpaired T-tests, were conducted to assess the differences between LLM and textbook responses.

Results A total of 90 surveys were completed. LLM responses were rated as more thorough ($p < 0.001$) but less concise ($p < 0.001$) than textbook responses. Textbooks were rated superior for answering questions on contraindications ($p = 0.027$) and complications ($p = 0.014$). ChatGPT was perceived as more accurate ($p = 0.018$), thorough ($p = 0.002$), and useful ($p = 0.026$) than Gemini. Gemini was rated lower in quality ($p = 0.30$) compared to ChatGPT along with being inferior to textbook answers for burn-related questions ($p = 0.017$) and anatomical questions ($p = 0.013$).

Conclusion While LLMs show promise in generating thorough educational content, they require improvement in conciseness, accuracy, and utility for practical case preparation. ChatGPT generally outperforms Gemini, indicating variability in LLM capabilities. Further development should focus on enhancing accuracy and consistency to establish LLMs as reliable tools in medical education and practice.

Keywords Artificial intelligence, Education, Plastic and reconstructive surgery, Medical school, Residency, Surgery, Chatgpt, Large language models, Gemini

*Correspondence:
Chandler Hinson
csh2121@jagmail.southalabama.edu
[1]Frederick P. Whiddon College of Medicine, University of South Alabama, 5851 USA North Drive, Mobile, AL 36688, USA
[2]Stanford Department of Surgery, Division of Plastic and Reconstructive Surgery, 770 Welch Road, Palo Alto, CA 94304, USA
[3]5795 USA North Drive, Mobile, AL 36608, USA

Introduction

In the realm of surgical education, the implementation of large language models (LLMs) marks a significant milestone. Large language models are a type of artificial intelligence algorithm that leverages deep learning techniques and large datasets to understand, summarize, and generate text-based content [1]. These models, such as OpenAI's Chat Generative Pre-Trained Transformer-4 (ChatGPT-4) and Google's Gemini, formally known as Bard, are reshaping the landscape of information acquisition by providing a quicker and more accessible gateway to necessary information [2, 3]. Unlike traditional sources, LLMs can process and generate human-like text based on vast amounts of data, making them powerful tools for a variety of applications, such as medical education. LLMs like GPT-4 and Gemini represent the forefront of artificial intelligence (AI) technology. The platforms have access to large amounts of data and utilizes sophisticated algorithms in understanding user requests and generating human-like responses. They can engage in detailed, nuanced conversations, provide summaries of extensive documents, and even generate creative content.

The integration of LLMs into surgical education has become a heavily studied area, with many papers stating the promising application for LLMs in surgical education [4–8]. The use of AI offers a dynamic, interactive experience that contrasts sharply with the static nature of textbooks. By allowing learners to pose specific questions, LLMs can provide easily accessible, immediate, tailored responses.

However, it is crucial to acknowledge the reservations surrounding LLMs. As a relatively novel technology, there are concerns about the accuracy of LLM responses along with the depth of knowledge the platforms' possess [9, 10]. While there has been extensive research of LLMs with regards to evidence-based applications in medical diagnostics, there is a level of skepticism about their reliability and effectiveness in medical education. Furthermore, issues such as potential biases in the training data, the inability to access proprietary medical databases, and the models' reliance on pre-existing internet data can lead to questions about the comprehensiveness and objectivity of the information it provides to learners [11–16].

This study aims to address these concerns by examining how LLMs respond to case questions for surgical trainees. This study analyzes the quality and efficiency of LLM responses and their utility in aiding surgical trainees' preparation, specifically for plastic and reconstructive surgery. By conducting surveys amongst medical students, research fellows, residents, and attendings, reviewers compared responses of LLMs with textbook question-answer pairs, the current gold-standard in surgical education and case preparation. We aim to discern strengths, limitations, and potential implications of integrating LLMs into plastic and reconstructive surgical education.

Methodology

Six common cases that trainees must be familiar with were selected for each major field of plastic and reconstructive surgery. The fields (and cases) included were craniofacial (cleft lip and palate), hand (carpal tunnel release), microsurgery (autologous breast reconstruction), burn (split-thickness skin graft), gender affirming (phalloplasty), and aesthetics (liposuction). Four types of questions were used to encompassed important areas of case preparation: clinical anatomy, indications of the procedure, contraindications of the procedure, and common complications. The list of questions is shown in appendix 1.

Based on the case, question-answer pairs were extracted from prevalent plastic surgery junior trainee case-preparation textbooks, namely *Essentials of Plastic Surgery (Second Edition)* and *Plastic & Reconstructive Surgery Board Review (Third Edition)* [17, 18]. Two leading AI LLMs, Gemini and ChatGPT-4, were also utilized to generate responses to each question, verbatim from those listed in the two previously mentioned case-preparation textbooks. No priming questions or prompts were employed in this study. Answers from these platforms were placed into surveys that were administered to reviewers of different surgical education levels. The LLMs question-answer pairs were compared directly to the answers provided in the traditional case-preparation textbooks.

The developed surveys were initially piloted amongst selected medical students, residents, and attendings. This study was granted IRB exempted status from the University of South Alabama. After the pilot, the surveys were distributed from October 1, 2023, to July 1, 2024, across multiple institutions to attendings, residents, research fellows, and medical students practicing or interested in plastic surgery. For the purposes of this study, "research fellows" refers specifically to medical students completing a dedicated research year between their third and fourth year of medical school, not post-graduate trainees in surgical residency or fellowship programs. Informed consent was obtained from all participants prior to enrolling in the study. Respondents assessed the textbook and AI responses using a 5-point Likert scale with 0 being the lowest/worst score and 5 being the highest/greatest score. Question-answer pairs were scored across the following categories: accuracy, thoroughness, usefulness for case preparation, brevity, and overall quality. Statistical analyses, including one-way ANOVA and unpaired T-tests, were performed to compare the performance

of LLMs versus textbook responses and evaluate the strength of responses for each question type or plastic surgery subspecialty. Statistical analysis was conducted in *STATA SE 16.0* with a p-value < 0.05 being designated as statistical significance.

Results

A total of 90 surveys were administered. Each clinical case had a total 15 responses. Table 1 shows the number of responses by surgical education levels.

Due to the limited sample size of each type of trainee, the authors were unable to stratify responses based on trainee type.

Textbook vs. LLMs

Compared to textbook responses, LLMs were viewed as more thorough ($p < 0.001$) and less concise ($p < 0.001$). Additionally, textbooks responses were seen as superior in providing information on contraindications ($p = 0.027$) and complications ($p = 0.014$) compared to LLMs. Table 2 shows the ANOVA outputs comparing LLMs and textbook views.

Textbook vs. ChatGPT

When comparing textbook answers strictly to ChatGPT outputs, ChatGPT responses were again viewed as being more thorough ($p < 0.001$) and less concise ($p < 0.001$). When stratifying by clinical case, there were no significant differences between textbook and ChatGPT responses. By question type, textbooks responses were seen as superior compared to ChatGPT ($p = 0.036$). Table 3 shows the unpaired T-Test outputs when comparing textbook and ChatGPT answers.

Textbook vs. Gemini

When comparing textbook answers to Gemini responses, Gemini responses were viewed as being less accurate ($p = 0.043$), more thorough ($p = 0.001$), less useful in case preparation ($p = 0.041$), and less concise ($p < 0.001$). When stratifying by clinical case, Gemini was perceived as being inferior in quality when answering questions related to burns ($p = 0.017$). When stratifying by question type, Gemini was also perceived as being inferior when providing answers related to anatomy ($p = 0.013$), contraindications ($p < 0.001$), and complications ($p < 0.001$). Table 4 shows the unpaired T-Test outputs when comparing textbook and Gemini answers.

ChatGPT vs. Gemini

When comparing ChatGPT directly to Gemini, ChatGPT is perceived as being more accurate ($p = 0.018$), more thorough ($p = 0.002$), more useful in case preparation ($p = 0.026$), and better in overall quality ($p = 0.03$). When stratifying by clinical case, ChatGPT was perceived as higher quality for cases within aesthetics ($p = 0.034$). ChatGPT was also perceived as being superior when answering questions about anatomy ($p = 0.016$) and contraindications ($p < 0.001$). Table 5 shows the unpaired T-Test outputs when comparing ChatGPT and Gemini answers.

Discussion

The findings of this study underscore both the potential benefits and challenges of integrating LLMs into medical education and clinical practice. The perceived thoroughness of LLM responses suggests that these models can serve as valuable tools for in-depth learning. However, the trade-off with brevity indicates a significant area for improvement, particularly for use in time-sensitive environments. This may negate the often-discussed benefit of a faster, compact, mobile resource in LLMs. While the thoroughness of LLM responses is commendable, practical application in clinical settings often requires concise and directly actionable information. The comprehensive detail provided by LLMs, while beneficial for educational purposes, can be overwhelming when quick decision-making is necessary. This suggests that LLMs need to be refined to strike a balance between providing detailed information and maintaining brevity to ensure they can be effective in both educational and clinical contexts.

The study also revealed differences in perceived accuracy and utility between ChatGPT and Gemini, with Gemini generally receiving lower ratings. These disparities highlight a critical area for improvement, particularly in ensuring the reliability of LLM-generated information. Future iterations of these models should focus on enhancing the accuracy and reliability of their responses, potentially through more rigorous training and validation processes that incorporate a broader range of medical data and scenarios.

Stratified analysis by clinical scenario further illuminates the variability in LLM performance. For instance, Gemini's perceived lower quality in addressing burn-related questions compared to textbook responses suggests that both platforms may be trained on different

Table 1 Frequency of responses by surgical education level by clinical case. GAS = Gender affirming surgery

	Craniofacial	GAS	Microsurgery	Hand	Burn	Aesthetic
Medical Students	3	4	6	4	3	3
Research Fellows	3	3	3	3	3	3
Residents	5	4	2	4	5	5
Attendings	4	4	4	4	4	4

Table 2 ANOVA results comparing views of LLM and textbook responses stratified by view of answer characteristic, clinical case, and question type. LLM = large Language models. GAS = Gender affirming surgery

Answer Characteristics					
		Mean	**Standard Deviation**	**F-Ratio**	***P*-Value**
Accuracy	Textbook	4.314	0.287	1.163	0.297
	LLM	4.109	0.416		
Thoroughness	Textbook	3.092	0.328	37.158	< 0.001*
	LLM	4.139	0.350		
Utility in Case Preparation	Textbook	3.794	0.274	1.238	0.282
	LLM	3.597	0.385		
Concise	Textbook	4.648	0.227	183.987	< 0.001*
	LLM	3.044	0.241		
Quality	Textbook	3.876	0.262	0.606	0.448
	LLM	3.726	0.429		
Clinical Case					
		Mean	**Standard Deviation**	**F-Ratio**	***P*-Value**
Craniofacial	Textbook	4.087	0.630	0.050	0.827
	LLM	4.023	0.459		
GAS	Textbook	3.570	0.539	0.258	0.620
	LLM	3.427	0.504		
Microsurgery	Textbook	3.871	0.585	0.384	0.546
	LLM	3.696	0.481		
Hand	Textbook	4.070	0.585	0.541	0.475
	LLM	3.860	0.484		
Burn	Textbook	4.283	0.448	2.569	0.133
	LLM	4.448	0.531		
Aesthetics	Textbook	3.787	0.766	0.647	0.436
	LLM	3.490	0.629		
Question Type					
		Mean	**Standard Deviation**	**F-Ratio**	***P*-Value**
Anatomy	Textbook	4.190	0.567	2.668	0.106
	LLM	3.848	0.894		
Indications	Textbook	3.869	0.652	0.196	0.659
	LLM	3.936	0.625		
Contraindications	Textbook	3.927	0.713	5.094	0.027*
	LLM	3.563	0.723		
Complications	Textbook	3.888	0.666	6.291	0.014*
	LLM	3.506	0.576		

*Statistically significant value

datasets and may lack uniform access to sub-specialty knowledge. This is also exemplified by the perception that Gemini's responses to aesthetic-related questions are inferior to those from ChatGPT. Ensuring consistency and reliability across all domains of medical knowledge is essential for these models to be trusted and widely adopted by healthcare professionals and trainees.

Importantly, this variability is not unique to plastic surgery. Similar concerns about accuracy, specialty-specific knowledge, and clinical applicability have been reported in other surgical specialties. For instance, in orthopedic surgery, LLMs like ChatGPT have demonstrated mixed performance, with acceptable explanations for straightforward topics but notable inaccuracies in more complex or nuanced clinical questions [19, 20]. In neurosurgery, while ChatGPT has been praised for its educational potential, it has also been shown to produce factual inaccuracies and lack subspecialty depth in procedural contexts [21]. These parallels suggest a broader trend: while LLMs have general utility across surgical domains, their performance may fall short when detailed specialty-specific knowledge is required.

These findings are also echoed in several non-surgical specialties, suggesting that some challenges and benefits of LLM integration may be broadly translatable across clinical disciplines. In internal medicine, for instance, LLMs have demonstrated strong performance on standardized exams such as the USMLE but have been criticized for producing hallucinations and lacking clinical nuance in complex patient scenarios (Kung et al., 2023)

Table 3 Unpaired T-Test outputs comparing textbook answers to ChatGPT outputs stratified by answer characteristics, clinical case, and question type. GAS = Gender affirming surgery

Answer Characteristics								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Accuracy	Textbook	4.314	0.287	0.687	0.415	-0.386, 0.265	0.146	0.687
	ChatGPT	4.370	0.214					
Thoroughness	Textbook	3.092	0.328	< 0.001	8.709	-1.649, -0.977	0.151	< 0.001*
	ChatGPT	4.405	0.170					
Utility in Case Preparation	Textbook	3.794	0.274	0.815	0.24	-0.382, 0.307	0.155	0.815
	ChatGPT	3.830	0.261					
Concise	Textbook	4.648	0.227	< 0.001	11.061	1.282, 1.928	0.145	< 0.001*
	ChatGPT	3.043	0.273					
Quality	Textbook	3.876	0.262	0.524	0.661	-0.467, 0.254	0.162	0.524
	ChatGPT	3.980	0.298					
Clinical Case								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Craniofacial	Textbook	4.087	0.630	0.825	0.2291	-0.885, 0.725	0.349	0.825
	ChatGPT	4.170	0.463					
GAS	Textbook	3.570	0.539	0.795	0.2686	-0.895, 0.708	0.347	0.795
	ChatGPT	3.660	0.559					
Microsurgery	Textbook	3.871	0.585	0.575	0.584	-0.663, 1.113	0.385	0.575
	ChatGPT	3.650	0.631					
Hand	Textbook	4.070	0.585	1	0	-0.822, 0.822	0.357	1
	ChatGPT	4.070	0.543					
Burn	Textbook	4.283	0.448	0.644	0.480	-0.596, 0.909	0.326	0.644
	ChatGPT	4.127	0.576					
Aesthetics	Textbook	3.787	0.766	0.815	0.2414	-1.0903, 0.884	0.428	0.815
	ChatGPT	3.890	0.574					
Question Type								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Anatomy	Textbook	4.190	0.567	0.892	0.136	-0.328, 0.287	0.154	0.892
	ChatGPT	4.211	0.715					
Indications	Textbook	3.869	0.652	0.100	1.668	-0.577, 0.052	0.158	0.100
	ChatGPT	4.131	0.667					
Contraindications	Textbook	3.924	0.718	0.975	0.031	-0.358, 0.369	0.182	0.975
	ChatGPT	3.918	0.803					
Complications	Textbook	3.888	0.666	0.036	2.141	0.023, 0.659	0.159	0.036*
	ChatGPT	3.547	0.666					

*Statistically significant value

[22]. In dermatology, studies have found that while ChatGPT can provide detailed information about common conditions, it lacks the diagnostic specificity required for more complex or rare skin diseases, reinforcing the need for human oversight [23]. Similarly, in psychiatry, LLMs have been explored for their potential in therapeutic dialogue generation and patient education, but concerns remain about their appropriateness, tone, and potential to reinforce harmful biases [24]. These parallels across non-surgical domains support the idea that while LLMs can augment education and preliminary clinical decision-making, domain-specific fine-tuning and careful integration remain essential for their safe and effective use.

User perception and trust are paramount in the adoption of LLMs in clinical practice. The study indicates that ChatGPT is perceived as more accurate and useful than Gemini, suggesting a higher level of trust in its responses. Building and maintaining this trust involves not only improving the technical performance of these models but also ensuring transparency in how LLMs source, process, and deliver information. Communicating their limitations and design clearly can help users better understand when and how to rely on their outputs.

The comprehensive nature of LLM responses can be a significant asset in medical education. These models can provide students with a rich source of information, facilitating a deeper understanding of complex medical concepts. However, educators should guide students on how to critically appraise and integrate LLM outputs with trusted resources. Encouraging thoughtful engagement

Table 4 Unpaired T-Test outputs comparing textbook answers to gemini outputs stratified by answer characteristics, clinical case, and question type. GAS = Gender affirming surgery

Answer Characteristics								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Accuracy	Textbook	4.314	0.287	0.043	2.319	0.018, 0.923	0.203	0.043*
	Gemini	3.843	0.406					
Thoroughness	Textbook	3.092	0.328	0.001	4.527	-1.164, -0.396	0.172	0.001*
	Gemini	3.872	0.266					
Utility in Case Preparation	Textbook	3.794	0.274	0.041	2.351	0.023, 0.839	0.183	0.041*
	Gemini	3.363	0.355					
Concise	Textbook	4.648	0.227	< 0.001	12.164	1.310, 1.900	0.132	< 0.001*
	Gemini	3.044	0.230					
Quality	Textbook	3.876	0.262	0.063	2.089	-0.027, 0.840	0.195	0.063
	Gemini	3.469	0.400					
Clinical Case								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Craniofacial	Textbook	4.087	0.630	0.569	0.594	-0.595, 1.009	0.348	0.569
	Gemini	3.880	0.456					
GAS	Textbook	3.570	0.539	0.221	1.326	-0.281, 1.041	0.287	0.221
	Gemini	3.190	0.346					
Microsurgery	Textbook	3.871	0.585	0.691	0.412	-0.574, 0.824	0.303	0.691
	Gemini	3.746	0.343					
Hand	Textbook	4.070	0.585	0.209	1.366	-0.287, 1.120	0.305	0.209
	Gemini	3.653	0.352					
Burn	Textbook	4.283	0.448	0.017	3.018	0.173, 1.294	0.243	0.017*
	Gemini	3.550	0.308					
Aesthetics	Textbook	3.787	0.766	0.109	1.805	-0.193, 1.587	0.386	0.109
	Gemini	3.090	0.399					
Question Type								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Anatomy	Textbook	4.190	0.567	0.013	2.553	0.105, 0.859	0.189	0.013*
	Gemini	3.709	0.962					
Indications	Textbook	3.869	0.652	0.970	0.038	-0.303, 0.292	0.149	0.970
	Gemini	3.874	0.595					
Contraindications	Textbook	3.924	0.718	< 0.001	4.394	0.360, 0.958	0.150	< 0.001*
	Gemini	3.265	0.521					
Complications	Textbook	3.888	0.666	< 0.001	3.566	0.213, 0.755	0.136	< 0.001*
	Gemini	3.404	0.448					

*Statistically significant value

with LLMs—rather than passive consumption—will be essential for fostering analytical skills among trainees.

Moreover, the specific strengths and weaknesses of ChatGPT and Gemini in different clinical scenarios suggest potential for targeted optimization. ChatGPT's strength in aesthetic-related questions and Gemini's relative weakness in burn-related questions indicate that fine-tuning LLMs to specific medical domains could enhance their relevance and reliability. This specialization may mirror the trajectory of LLMs in other surgical fields, where custom training on subspecialty datasets has been proposed as a way to overcome generalization limitations [25, 26].

Despite the valuable insights provided by this study, several limitations must be acknowledged. First, the small sample size limited our ability to explore differences in perceptions across training levels—such as between medical students, research fellows, residents, and attendings. This limitation has important implications, as individuals at different stages of training may possess varying levels of clinical knowledge and familiarity with surgical decision-making. Medical students, in particular, may not yet have the depth of clinical experience to critically evaluate the accuracy or clinical relevance of LLM-generated responses. As a result, their assessments may differ from those of more advanced trainees or practicing surgeons, potentially introducing bias into perceived ratings of accuracy, utility, or quality.

Survey fatigue may have also affected the reliability of responses, as the survey's length and cognitive demand

Table 5 Unpaired T-Test outputs comparing ChatGPT answers to Gemini outputs stratified by answer characteristics, clinical case, and question type. GAS = Gender Affirming Surgery

Answer Characteristics								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Accuracy	ChatGPT	4.370	0.214	0.018	2.835	0.114, 0.949	0.187	0.018*
	Gemini	3.843	0.406					
Thoroughness	ChatGPT	4.405	0.170	0.002	4.134	0.245, 0.820	0.129	0.002*
	Gemini	3.872	0.266					
Utility in Case Preparation	ChatGPT	3.830	0.261	0.026	2.601	0.067, 0.869	0.180	0.026*
	Gemini	3.363	0.355					
Concise	ChatGPT	3.043	0.273	0.995	0.007	-0.326, 0.324	0.146	0.995
	Gemini	3.044	0.230					
Quality	ChatGPT	3.980	0.298	0.030	2.529	0.061, 0.966	0.203	0.030*
	Gemini	3.469	0.400					
Clinical Case								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Craniofacial	ChatGPT	4.170	0.463	0.353	0.987	-0.383, 0.957	0.291	0.353
	Gemini	3.880	0.456					
GAS	ChatGPT	3.660	0.559	0.146	1.610	-0.205, 1.151	0.294	0.146
	Gemini	3.190	0.346					
Microsurgery	ChatGPT	3.650	0.631	0.764	0.311	-0.841, 0.641	0.321	0.764
	Gemini	3.746	0.343					
Hand	ChatGPT	4.070	0.543	0.188	1.44	-0.250, 1.084	0.289	0.188
	Gemini	3.653	0.352					
Burn	ChatGPT	4.127	0.576	0.084	1.974	-0.097, 1.250	0.292	0.084
	Gemini	3.550	0.308					
Aesthetics	ChatGPT	3.890	0.574	0.034	2.56	0.079, 1.521	0.313	0.034*
	Gemini	3.090	0.399					
Question Type								
		Mean	**Standard Deviation**	**Two-Tailed**	**T-Value**	**95% Confidence Interval**	**Standard Error**	***P*-Value**
Anatomy	ChatGPT	4.211	0.715	0.016	2.482	0.099, 0.907	0.203	0.016*
	Gemini	3.709	0.962					
Indications	ChatGPT	4.131	0.667	0.093	1.702	-0.044, 0.559	0.151	0.093
	Gemini	3.874	0.595					
Contraindications	ChatGPT	3.918	0.803	< 0.001	4.036	0.330, 0.976	0.162	< 0.001*
	Gemini	3.265	0.521					
Complications	ChatGPT	3.547	0.666	0.296	1.052	-0.128, 0.414	0.136	0.296
	Gemini	3.404	0.448					

*Statistically significant value

may have decreased participant engagement over time. Additionally, the subjective nature of Likert-scale ratings and the possibility of individual biases toward or against technology introduce variability, as user perceptions may not always align with objective measures of LLM performance. The limited scope of clinical scenarios and question types examined further restricts the generalizability of our findings, as the full breadth and complexity of plastic surgery practice were not fully captured.

Addressing these limitations through larger and more diverse samples, minimizing survey fatigue, incorporating objective performance assessments, and expanding the clinical breadth of the questionnaire will be essential in future work. In addition, we propose the use of prompt priming—for example, instructing the LLM to respond as a plastic surgery expert communicating with a medical colleague—to potentially improve the quality and relevance of generated responses. Further research is needed to evaluate whether such strategies enhance the educational and clinical utility of LLMs for subspecialty training.

Conclusion

While ChatGPT and Gemini offer promising capabilities in generating thorough and detailed responses, there are significant areas for improvement, particularly in conciseness, accuracy, and utility for practical case preparation. ChatGPT generally outperforms Gemini, suggesting that different LLMs may have varying strengths and weaknesses. Future developments should aim to enhance

the balance between detail and brevity, improve accuracy, and ensure consistent performance across different clinical scenarios and question types. These advancements will be critical in establishing LLMs as reliable and effective tools in both medical education and clinical practice. As these models evolve, their integration into healthcare settings holds the potential to transform medical education and support healthcare professionals in delivering high-quality patient care.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-07550-8.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
CH: Study Design, Data Collection, Data Analysis, Manuscript; CS: Study Design, Data Collection, Data Analysis, Manuscript; RN: Study Design, Data Analysis, Manuscript.

Funding
The study received no financial support from any external funding source. No authors were supported by any grants or funding sources.

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
IRB approval (expedited) was granted from the University of South Alabama. Informed consent was collected from each participant prior to partaking in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Clinical trial number
Not applicable.

Received: 11 September 2024 / Accepted: 16 June 2025
Published online: 01 July 2025

References

1. Amazon Web Solutions. What are large language models? - LLM AI explained - AWS. Amazon Web Services, Inc. Accessed July 21. 2024. https://aws.amazon.com/what-is/large-language-model/
2. OpenAi. OpenAI - ChatGPT. 2024. Accessed July 21, 2024. https://openai.com/research/
3. Google. Gemini. Gemini. 2024. Accessed July 21, 2024. https://gemini.google.com
4. Kirubarajan A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. J Surg Educ. 2022;79(2):500–15. https://doi.org/10.1016/j.jsurg.2021.09.012
5. Guerrero DT, Asaad M, Rajesh A, Hassan A, Butler CE. Advancing surgical education: the use of artificial intelligence in surgical training. Am Surgeon™. 2023;89(1):49–54. https://doi.org/10.1177/00031348221101503
6. Bilgic E, Gorgy A, Yang A, et al. Exploring the roles of artificial intelligence in surgical education: a scoping review. Am J Surg. 2022;224(1, Part A):205–16. https://doi.org/10.1016/j.amjsurg.2021.11.023
7. Satapathy P, Hermis AH, Rustagi S, Pradhan KB, Padhi BK, Sah R. Artificial intelligence in surgical education and training: opportunities, challenges, and ethical considerations– correspondence. Int J Surg. 2023;109(5):1543. https://doi.org/10.1097/JS9.0000000000000387
8. Vedula SS, Ghazi A, Collins JW, et al. Artificial intelligence methods and artificial intelligence-enabled metrics for surgical education: a multidisciplinary consensus. J Am Coll Surg. 2022;234(6):1181. https://doi.org/10.1097/XCS.0000000000000190
9. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg. 2024;110(6):3701. https://doi.org/10.1097/JS9.0000000000001312
10. Gradon KT. Generative artificial intelligence and medical disinformation. BMJ. 2024;384:q579. https://doi.org/10.1136/bmj.q579
11. Nguyen T. ChatGPT in medical education: a precursor for automation bias?? JMIR Med Educ. 2024;10(1):e50174. https://doi.org/10.2196/50174
12. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large language models (LLMs). Npj Digit Med. 2024;7(1):1–11. https://doi.org/10.1038/s41746-024-01157-x
13. Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. Nat Mach Intell. 2021;3(6):461–3. https://doi.org/10.1038/s42256-021-00359-2
14. Yeung JA, Kraljevic Z, Luintel A et al. AI chatbots not yet ready for clinical use. Published online March 20, 2023:2023.03.02.23286705. https://doi.org/10.1101/2023.03.02.23286705
15. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate Racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1):e12–22. https://doi.org/10.1016/S2589-7500(23)00225-X
16. Liu Z, Zhang L, Wu Z, et al. Surviving ChatGPT in healthcare. Front Radiol. 2024;3:1224682. https://doi.org/10.3389/fradi.2023.1224682
17. Janis JE, editor. Essentials of plastic surgery. 2nd ed. CRC; 2015. https://doi.org/10.1201/b16610
18. Lin SJ, Hijjawi JB. *Plastic and reconstructive surgery board review: pearls of wisdom, third edition*. 3rd edition. McGraw Hill / Medical; 2016.
19. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT performance on the orthopaedic in-training examination. JB JS Open Access. 2023;8(3):e23.00056.
20. Jain N, Gottlich C, Fisher J, Campano D, Winston T. Assessing chatgpt's orthopedic in-service training exam performance and applicability in the field. J Orthop Surg Res. 2024;19(1):27. https://doi.org/10.1186/s13018-023-04467-0
21. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. 2023;93(6):1353–65. https://doi.org/10.1227/neu.0000000000002632
22. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
23. Goktas P, Grzybowski A. Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review. J Clin Med. 2024;13(19):5909. https://doi.org/10.3390/jcm13195909
24. Cheng S, Chang C, Chang W, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci. 2023;77(11):592–6. https://doi.org/10.1111/pcn.13588
25. Khalpey Z, Kumar U, King N, Abraham A, Khalpey AH. Large language models take on cardiothoracic surgery: a comparative analysis of the performance of four models on American Board of Thoracic Surgery Exam Questions in 2023. Cureus. 16(7):e65083. https://doi.org/10.7759/cureus.65083
26. Long C, Subburam D, Lowe K et al. ChatENT: augmented large language models for expert knowledge retrieval in otolaryngology - head and neck surgery. Published online August 21, 2023:2023.08.18.23294283. https://doi.org/10.1101/2023.08.18.23294283

Publisher's note