

Research

Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study

Ishith Seth, BBiomed(Hons), MD, MSurg[®]; Bryan Lim, MD; Yi Xie, LLB, BA, MD; Jevan Cevik, BBiomed, MD; Warren M. Rozen, MBBS, BMSc, MD, PhD, FRACS; Richard J. Ross, BSc(Hons), MBBS, PhD, FRACS; and Mathew Lee, MBBS, FRACS

Aesthetic Surgery Journal Open Forum 2023, 1–9

© The Author(s) 2023. Published by Oxford University Press on behalf of The Aesthetic Society.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

<https://doi.org/10.1093/asjof/ojad084>
www.asjopenforum.com

OXFORD
UNIVERSITY PRESS

Abstract

Background: Large language models (LLMs) are emerging artificial intelligence (AI) technologies refining research and healthcare. However, the impact of these models on presurgical planning and education remains under-explored.

Objectives: This study aims to assess 3 prominent LLMs—Google’s AI BARD (Mountain View, CA), Bing AI (Microsoft, Redmond, WA), and ChatGPT-3.5 (Open AI, San Francisco, CA) in providing safe medical information for rhinoplasty.

Methods: Six questions regarding rhinoplasty were prompted to ChatGPT, BARD, and Bing AI. A Likert scale was used to evaluate these responses by a panel of Specialist Plastic and Reconstructive Surgeons with extensive experience in rhinoplasty. To measure reliability, the Flesch Reading Ease Score, the Flesch–Kincaid Grade Level, and the Coleman–Liau Index were used. The modified DISCERN score was chosen as the criterion for assessing suitability and reliability. A *t* test was performed to calculate the difference between the LLMs, and a double-sided *P*-value <.05 was considered statistically significant.

Results: In terms of reliability, BARD and ChatGPT demonstrated a significantly (*P* < .05) greater Flesch Reading Ease Score of 47.47 (±15.32) and 37.68 (±12.96), Flesch–Kincaid Grade Level of 9.7 (±3.12) and 10.15 (±1.84), and a Coleman–Liau Index of 10.83 (±2.14) and 12.17 (±1.17) than Bing AI. In terms of suitability, BARD (46.3 ± 2.8) demonstrated a significantly greater DISCERN score than ChatGPT and Bing AI. In terms of Likert score, ChatGPT and BARD demonstrated similar scores and were greater than Bing AI.

Conclusions: BARD delivered the most succinct and comprehensible information, followed by ChatGPT and Bing AI. Although these models demonstrate potential, challenges regarding their depth and specificity remain. Therefore, future research should aim to augment LLM performance through the integration of specialized databases and expert knowledge, while also refining their algorithms.

Level of Evidence: 5

Editorial Decision date: September 8, 2023; online publish-ahead-of-print September 14, 2023.



Dr Seth and Prof Rozen are research staff members, and Mr Lim is a medical student, Monash University, Melbourne, Australia. Dr Xie is a plastic surgery registrar, Dr Cevik is a plastic surgery resident, and Drs Ross and Lee are specialist plastic surgeons at Peninsula Health, Melbourne, Victoria, Australia.

Corresponding Author:

Dr Ishith Seth, Faculty of Medicine, Monash University, Melbourne, Victoria 3004, Australia.

E-mail: ishithseth1@gmail.com

Rhinoplasty remains one of the most sought-after surgical procedures across the globe.¹ In 2018, the United States saw over 200,000 such operations, making it the third most common plastic surgery performed in the nation.² The nose is the central focal point of the face, with its proportions and symmetry having a significant impact on overall facial aesthetics.¹

With the recent advancements in artificial intelligence (AI) and natural language processing, large language models (LLMs) have demonstrated remarkable capabilities in the surgical field.³⁻⁷ The transformative power of AI has pervaded a plethora of domains, including healthcare, where it has revolutionized diagnostics, treatment planning, and patient care.⁸ In recent times, AI-driven LLMs, such as ChatGPT (Open AI, San Francisco, CA), BARD (Google, Mountain View, CA), and Bing AI (Microsoft, Redmond, WA), have gained considerable attention for their ability to comprehend and generate human-like text based on massive amounts of training data.^{4,9,10} The integration of AI techniques holds tremendous potential to reshape the field of plastic surgery by providing insights into patient-specific rhinoplasty outcomes, surgical planning, and postoperative care. Moreover, the utilization of LLMs can facilitate efficient extraction and analysis of scientific literature, enhancing the understanding of best practices and potential complications associated with rhinoplasty.¹¹ However, alongside the potential benefits, it is imperative to address the challenges and limitations associated with deploying AI-driven language models in a medical context, to ensure the reliability, accuracy, and ethical use of generated information.

This study aims to investigate the efficacy of employing LLMs in obtaining and synthesizing information about rhinoplasty. We assess the mechanisms driving these models to appraise their capacity to augment presurgical planning or medical decision making. Additionally, we scrutinize the ethical and pragmatic aspects of incorporating AI-powered language models into clinical settings, highlighting potential advantages and challenges.

METHODS

We engaged ChatGPT-3.5, BARD, and Bing AI with a series of 6 questions targeting various technical aspects of rhinoplasty. These queries were devised by 3 board-certified plastic surgeons who were fellows of the Royal Australasian College of Surgeons with over 25 years of combined experience performing Rhinoplasty and expertise in facial reconstructive surgery. The aim was to evaluate the breadth and depth of the LLMs' knowledge of rhinoplasty. The accuracy of responses was determined by comparing them with current healthcare guidelines for rhinoplasty and through evaluation by the panel of plastic surgeons through a Likert scale (Table 1). If any differences

in the Likert scale arose, these were discussed until consensus was achieved. During the process, the LLMs were requested to supply 5 high-quality references to substantiate their responses. The references and literature were then assessed for relevance and quality and compared to our preliminary database searches on PubMed (National Institutes of Health, Bethesda, MD), Web of Science (Clarivate Analytics, London, UK), Scopus (Elsevier, Amsterdam, the Netherlands), EMBASE (Elsevier), Cochrane CENTRAL (Wiley, Hoboken, NJ), and Google search. Lastly, the LLM responses were quantitatively analyzed using validated tools (Table 2). The LLMs were requested to limit their answers to 200 words. To standardize response comparisons and mimic real-world preference for brevity, this has been validated in previous LLM studies.^{7,12} To measure the reliability of the response, the Flesch Reading Ease Score, the Flesch–Kincaid Grade Level, and the Coleman–Liau Index were utilized. The Flesch Reading Ease Score, on a scale of 0 to 100, indicates the readability of the text—a higher score translates to easier comprehension. The Flesch–Kincaid Grade Level and the Coleman–Liau Index, with scales ranging from 0 to ∞ , reflect the complexity of the text and the education level required for understanding it, respectively, and in both instances, a higher score implies more complexity. We also employed the modified DISCERN score, which ranges from 16 to 80. A higher score on this scale represents superior quality and more balanced treatment option information. This score was used to assess the appropriateness of the response, thereby ensuring the provided information's quality and relevance for patients. *T* test was employed to compare the differences between the 3 LLMs, and a *P*-value $<.05$ was considered statistically significant.

The inclusion criteria for this study encompassed the initial response given by the LLMs to each question. No exclusion criteria were applied to the answers generated. Since this study entailed an observational case study of a publicly available AI LLM, no institutional ethical approval was required.

RESULTS

In response to the first inquiry in the Figure, ChatGPT accurately delineated surgical interventions for internal nasal valve dysfunction, emphasizing the importance of enhancing cross-sectional area but overlooked discussing the angle between upper lateral cartilages and the anterior septal edge. Among 5 referenced sources, only 2 were found in the literature,^{13,14} 1 being contextually relevant.¹³ BARD provided a contextual overview, briefly discussing the mechanism and potential side effects, before identifying spreader, Alar batten, and Butterfly grafts,^{15,16} but failed to discuss the spreader graft's limitations in expanding a collapsed lateral nasal wall as noted by Teymoortash

Table 1. Evaluation of the Reponses of Large Language Model Platforms

Criteria	ChatGPT (Open AI)	Bing AI (Microsoft)	Google's BARD
The large language model provides accurate answers to questions.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model is reliable when generating factual information.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model is proficient at understanding complex questions and providing appropriate answers.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input checked="" type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model provides comprehensive information when answering questions.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model generates content that covers all relevant aspects of a subject.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model is able to provide in-depth information for a wide range of topics.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model is a valuable source of general knowledge.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input checked="" type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model is well-versed in a variety of subjects.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input checked="" type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model can provide useful insights and perspectives on various topics.	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input checked="" type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model rarely makes errors when referencing sources.	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree
The large language model is consistent in providing accurate citations.	<input type="checkbox"/> 1—Strongly disagree <input checked="" type="checkbox"/> 2—Disagree <input type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree	<input type="checkbox"/> 1—Strongly disagree <input type="checkbox"/> 2—Disagree <input checked="" type="checkbox"/> 3—Neither agree or disagree <input type="checkbox"/> 4—Agree <input type="checkbox"/> 5—Strongly agree

et al. Only 1 of its 5 references could be corroborated in the literature. Bing AI listed the same techniques as BARD, providing a brief historical context for spreader grafts and the primary etiology of internal nasal valve dysfunction. It produced legitimate references, yet only 2 were academic.

Assessing the responses, ChatGPT presented a broader range of surgical options, including flaring sutures, as a potential alternative to grafts.

ChatGPT's response to the second inquiry in [Supplemental Figure 1](#) followed a similar pattern as the first.

It discussed fortifying the lateral nasal wall but ignored the significance of the lateral crus and nasal septum in external nasal valve function. It correctly listed methods for nasal tip deprojection, alar rim reinforcement, and crura realignment,¹⁷ although its 5 references proved to be false. BARD accurately outlined external valve collapse, its dysfunction, and its causes, followed by a brief discussion of 3 surgical interventions. Like ChatGPT, all its references were erroneous. Bing AI's shorter response focused on the general concept of deprojecting the over-projected nose, but only hinted at the use of lateral crura strut and alar grafts, failing to expand further. None of its references were academic in nature. Overall, ChatGPT provided the most detailed and diverse surgical options for the correction of external valve collapse.

Supplemental Figure 2 shows the LLMs' management for caudal septal dislocation. ChatGPT accurately mentioned the tongue-in-groove technique,¹⁸ but misattributed techniques such as medial crural overlay, caudal septal extension, and transdermal suture to septal dislocation correction, which are typically adjuncts for nasal tip projection correction, and not primary methods for caudal septal dislocation. Only 1 of its references could be found in the literature, but it related to nasal tip contouring,¹⁹ not supporting ChatGPT's response. BARD focused on surgical techniques, but used vague terminologies such as "Suturing," "Grafting," and "Septoplasty." All its references were incorrectly cited, missing author names, journal volumes, and page numbers. Bing AI was again brief, mainly discussing septoplasty and vaguely discussing open approaches and grafts. It was able to provide 1 academic article, matching ChatGPT in this response. Consequently, ChatGPT outperformed both by offering more detailed and numerous surgical options.

The response provided by ChatGPT to the fourth prompt in Supplemental Figure 3 exhibited comparable structure to its previous responses. It appropriately emphasized reducing tissue size to improve nasal airflow but overlooked the potential impacts on mucociliary clearance and local immune function. It provided a list of surgical techniques for treating this condition,^{20,21} although its references were again unverifiable. BARD underscored the pathophysiology and symptoms of turbinate hypertrophy and discussed 2 surgical corrections. Analogous to its previous response, it omitted crucial citation details like author names and journal volumes, while providing a link to a specialist's website, resulting in low-quality references. Bing AI offered a larger array of options, including nonsurgical ones such as nasal sprays and antihistamines, but only suggested inferior turbinate resection as a surgical option.²² It marginally outperformed ChatGPT and BARD in this instance by supplying 2 academic articles.

In ChatGPT's response to the fifth prompt in Supplemental Figure 4. The analysis exhibited a relatively restricted purview. It covered nasal tip support in

rhinoplasty but overlooked the complex relationship between the junction of the medial and the lateral crurae of each lower lateral cartilage and its effect on tip support.^{23,24} Unfortunately, ChatGPT's references were spurious and could not be verified. BARD provided a brief background on submucous resection's impact on tip support and discussed 3 corrective surgical methods. All of its references were absent in the literature. Bing AI failed to provide any relevant information or recommendations.

The objective of the sixth prompt in Supplemental Figure 5 was to assess ChatGPT's recommendations regarding the timing of osteotomy performance during rhinoplasty. While the response accurately identified 1 rationale for conducting an osteotomy during rhinoplasty, it omitted others, such as addressing open roof deformities.²⁵ Although the open and closed approaches were discussed, the response failed to delineate the specific circumstances warranting their application, instead focusing on procedural technicalities. All 5 references provided were truncated and, as such, could not be further analyzed. Regrettably, both BARD and Bing AI erroneously interpreted the question, proposing surgical interventions for fractured nasal bones.

When comparing the 3 LLMs for readability and reliability, BARD consistently outperformed, registering the highest Flesch Reading Ease Score (47.5 ± 15.32), Flesch-Kincaid Grade Level (9.71 ± 3.12), and DISCERN score (46.3 ± 2.88), indicative of superior comprehensibility and medical advice alignment with clinical guidelines. This was followed by ChatGPT, and then Bing AI, which lagged in all categories. The only comparisons that yielded statistically nonsignificant outcomes ($P > .05$) were those between ChatGPT and BARD regarding their readability indices and between ChatGPT and Bing AI when assessing the DISCERN scores. All other comparative analyses demonstrated statistical significance ($P < .05$).

DISCUSSION

LLMs have ascended to prominence in the medical sector owing to their proficient rapid information retrieval and algorithmic decision-making capabilities.²⁶ The integration of AI into the planning of rhinoplasty operations has seen remarkable growth,^{27,28} particularly in the context of the burgeoning utilization of LLMs in healthcare. Given this trajectory, this research study comparatively appraises the potential of Google's BARD, Bing AI, and ChatGPT in enhancing preoperative planning and decision making.

It is worth noting that the LLMs were not intended to serve as a literature search or reference engine. ChatGPT and BARD, machine-learning models trained on diverse internet data, generate text based on statistical models, covering a wide array of topics. However, they cannot source literature to support their content, despite their predictive text generation creating an illusion of authenticity as

Table 2. Readability and Reliability of the Responses of Large Language Models

Model	Prompts	Readability			Suitability
		Flesch reading ease score	Flesch–Kincaid grade level	Coleman–Liau index	DISCERN score
ChatGPT (Open AI)	Correcting internal valve dysfunction	42.98	9.56	15	42
	Correcting external valve collapse	41.43	9.69	16	42
	Correcting caudal septal dislocation	33.97	10.59	6	44
	Managing turbinate hypertrophy	13.27	13.62	12	41
	Managing tip support after submucous resection	49.02	8.46	18	43
	When to do nasal bone fractures in rhinoplasty	45.42	8.99	12.17	41
Mean (SD)		37.68 (\pm 12.96)	10.15 (\pm 1.84)	12.00 (5.10)	42.17 (\pm 1.17)
Google's BARD	Correcting internal valve dysfunction	61.26	6.98	10	44
	Correcting external valve collapse	61.63	7.25	9	47
	Correcting caudal septal dislocation	36.08	11.27	13	42
	Managing turbinate hypertrophy	23.03	15.2	14	48
	Managing tip support after submucous resection	54.89	8.15	9	47
	When to do nasal bone fractures in rhinoplasty	47.91	9.4	10	50
Mean (SD)		47.47 (\pm 15.32)	9.71 (\pm 3.12)	10.83 (\pm 2.14)	46.33 (\pm 2.88)
Bing AI(Microsoft)	Correcting internal valve dysfunction	25.95	20	16	39
	Correcting external valve collapse	10.17	20.86	14	30
	Correcting caudal septal dislocation	14.41	19.2	14	36
	Managing turbinate hypertrophy	4.89	18.37	19	38
	Managing tip support after submucous resection	NA	25.34	14	NA
	When to do nasal bone fractures in rhinoplasty	36.03	13.56	12	46
Mean (SD)		18.29 (\pm 12.59)	18.25 (\pm 4.84)	12.00 (5.10)	35.75 (\pm 4.03)

NA, not applicable; SD, standard deviation.

observed by Saleem et al.²⁹ Bing AI, benefiting from real-time internet access, does not share this limitation, but its references are subpar, offering few academic articles, and failing to cite them appropriately. Accordingly, authoring scientific literature still necessitates the guidance and supervision of experts.³⁰

ChatGPT demonstrates promise in enhancing preoperative planning in rhinoplasty procedures. Open AI is currently working on a feature that enables ChatGPT to analyze and interpret images, which, when combined with machine-learning algorithms and perioperative photographic patient records, could yield highly accurate predictions of

rhinoplasty outcomes. Google also recently announced a forthcoming update to BARD, incorporating image interpretation, including those of medical relevance. This enhancement has significant potential to bolster the presurgical planning phase for rhinoplasties. Such predictions could provide surgeons with innovative ways to plan procedures and identify potential complications.³¹

However, certain limitations within its current algorithm need to be addressed. As shown in the [Figure](#) and [Supplemental Figures 1 to 5](#), all 3 LLMs generated potentially incomplete and unreliable lists of surgical management, omitting feasible surgical interventions. Moreover,

ChatGPT

In 200 words, how do you correct internal valve dysfunction in rhinoplasty? List 5 relevant references.

Internal valve dysfunction in rhinoplasty can be corrected through various surgical techniques that aim to increase the cross-sectional area of the internal nasal valve. The most common techniques include spreader grafts, flaring sutures, butterfly grafts, batten grafts, and alar batten grafts.

Spreader grafts are used to widen the middle part of the nose by inserting cartilage between the septum and the upper lateral cartilages. Flaring sutures are placed between the lower lateral cartilages and the septum to pull the cartilages outward and open up the nasal airway. Butterfly grafts, batten grafts, and alar batten grafts are used to strengthen the lateral nasal wall and prevent collapse of the internal nasal valve during inspiration.

The choice of technique depends on the individual patient's anatomy and the extent of internal valve dysfunction. A thorough preoperative assessment is essential to determine the appropriate surgical approach.

Here are five relevant references that discuss the correction of internal valve dysfunction in rhinoplasty:

- Ahmad J, Ikram M, Ashraf S. Correction of Internal Nasal Valve Dysfunction with Spreader Grafts in Septorhinoplasty. *Annals of King Edward Medical University*. 2020;26(2):132-137.
- Gruber RP, Park E, Newman J. Internal Nasal Valve Repair: A Comparison of Flaring Sutures and Spreader Grafts. *Plastic and Reconstructive Surgery*. 2007;119(2):839-845.
- Toriumi DM, Josen J, Weinberger M, Tardy ME Jr. Use of alar batten grafts for correction of nasal valve collapse. *Arch Otolaryngol Head Neck Surg*. 1997;123(8):802-808.
- Pasha R, Saleem M, Yaqoob N, Muhammad N. A Comparative Study on the Use of Spreader Grafts and Butterfly Grafts for Correction of Internal Nasal Valve Dysfunction. *Journal of Rawalpindi Medical College (JRMCI)*. 2021;25(1):30-33.
- Most SP. Anterior Septal Reconstruction: Outcomes After a Modified Extracorporeal Septoplasty Technique. *Archives of Facial Plastic Surgery*. 2006;8(3):202-207.

Google's BARD

In 200 words, how do you correct internal valve dysfunction in rhinoplasty? List 5 relevant references.

Internal valve dysfunction (IVD) is a common problem that can occur after rhinoplasty. It is caused by a narrowing of the internal nasal valve, which is the area where the nasal septum meets the lateral nasal walls. This narrowing can lead to difficulty breathing through the nose.

There are a number of surgical techniques that can be used to correct IVD. The most common technique is to place spreader grafts. Spreader grafts are small pieces of cartilage that are placed between the upper lateral cartilages and the septum. This helps to widen the internal nasal valve and improve airflow.

Other techniques that can be used to correct IVD include:

- Alar batten grafts: These are small pieces of cartilage that are placed along the lateral nasal walls. This helps to prevent the lateral nasal walls from collapsing inward and narrowing the internal nasal valve.
- Butterfly grafts: These are small pieces of cartilage that are shaped like butterflies. They are placed in the space between the upper lateral cartilages and the septum. This helps to widen the internal nasal valve and improve airflow.

The best technique for correcting IVD depends on the individual patient's anatomy and the severity of the problem.

Here are 5 relevant references:

- Sheen JH, Sheen AP. Spreader grafts: a method of widening the nasal valve. *Plast Reconstr Surg*. 1984;73(5):597-601.
- Gunter JP, Rohrich RJ, Friedman RM. The use of spreader grafts to correct internal nasal valve collapse. *Arch Otolaryngol Head Neck Surg*. 1997;123(1):85-90.
- Rohrich RJ, Hollier LH, Muzaffar AR, et al. The role of the lateral crural strut graft in rhinoplasty. *Plast Reconstr Surg*. 2002;109(6):1889-1902.
- Gunter JP, Tebbetts JB. Alar batten grafts: a new technique for correcting internal nasal valve collapse. *Plast Reconstr Surg*. 2001;107(6):1431-1437.
- Constantian MB. The internal nasal valve: anatomy, physiology, and surgical considerations. *Facial Plast Surg*. 2004;20(1):1-13.

Bing's AI

In 200 words, how do you correct internal valve dysfunction in rhinoplasty? List 5 relevant references.

Searching for: correct internal valve dysfunction rhinoplasty

Generating answers for you...

According to a study on PubMed, internal valve dysfunction is the most common site of obstruction in rhinoplasty. Its correction often involves alar batten grafts and reconstruction of the lateral crura¹. Many surgical techniques have been developed to correct nasal valve pathology according to the place of maximum collapse. In 1984, Sheen described the use of spreader grafts to correct the internal nasal valve by widening the nasal middle vault after reduction rhinoplasty². In addition, batten grafts are useful techniques for correction of nasal valve collapse in patients that haven't undergone rhinoplasty³. Internal nasal valve dysfunction is a frequent complication secondary to old school destructive rhinoplasty. The principal cause is over-resection of the lateral cartilages during hump removal⁴.

Here are 5 relevant references:

- Functional considerations in revision rhinoplasty - PubMed
- Management of the Internal Nasal Valve - Rhinoplasty Archive
- Management of Common Complications in Rhinoplasty and Medical ...
- Internal Valve Stenosis Rhinoplasty Treatment & Management - Medscape
- Internal Valve Stenosis Rhinoplasty - Medscape

I hope this helps!

Learn more: 1. pubmed.ncbi.nlm.nih.gov 2. rhinoplastyarchive.com +3 more 1 of 30

Figure. In 200 words, how do you correct internal valve dysfunction in rhinoplasty? List 5 relevant references.

in Supplemental Figure 5, BARD and Bing AI deviated from the prompt, dedicating much of their response to explaining surgical procedures rather than concentrating on when they should be performed. Additionally, as evidenced in Supplemental Figure 2, certain surgical techniques were inappropriately suggested by ChatGPT for the given scenario. One possible approach to addressing these limitations is refining the prompt, as highlighted by Li et al, who emphasized the importance of properly crafting prompts.³² Conversely, before integrating LLMs into medical practice, it is crucial to rectify these algorithmic shortcomings to ensure reliable and accurate information.

BARD demonstrated superior readability and reliability than ChatGPT and Bing AI, making it ideal for nonmedical users and as a preliminary medical advisory tool in medically underserved areas. ChatGPT's responses, although understandable, were less comprehensive partly due to its failure to contextualize the defect and emphasize the need for repair as BARD does. This is crucial as patients may retrieve knowledge from these LLMs and other sources, empowering them to be actively involved in presurgery

decision making and planning. Despite Bing AI's complex medical language, it often fell short of a detailed explanation. The study by Zhu et al supports the impact of different training data and preprocessing methods on LLMs' readability,³³ suggesting further research could optimize algorithms and improve comprehensibility. Furthermore, the *t* test demonstrated that most comparisons were statistically significant. However, the comparisons pertaining to readability between ChatGPT and BARD, as well as the DISCERN score between ChatGPT and Bing AI, were statistically insignificant ($P > .05$). This implies that ChatGPT may not necessarily be less readable than BARD and more reliable than Bing AI. Hence, the authors agree with the current literature and suggest further studies comparing these LLMs on these parameters to obtain more reliable and statistically significant results.

Concerningly, none of the models assessed the benefits, risks, and postoperative quality of life related to the suggested procedures, which are vital for presurgical planning. Rhinoplasties, for example, can significantly impact nasal functionality and a patient's psychological state

postsurgery.³⁴ It would have been beneficial if the models had briefly touched on measures such as the Health Measurement Questionnaire and Glasgow Benefit Inventory.³⁵ The authors recommend fostering collaborations between AI developers and clinical experts to improve these LLMs' performance. Using specialized databases and expert knowledge could enhance the models' accuracy and depth of information. Moreover, ensuring the traceability and credibility of AI-generated content is key to building users' trust and accountability.

ChatGPT, Google's BARD, and Bing AI are revolutionizing the field with their advanced contextual understanding and predictive capabilities.³⁶ Integrated into various applications, they show promise in enhancing preoperative planning, decision making, and patient education. By synthesizing evidence-based recommendations, they exhibit the potential to equip surgeons with the latest best practices. Through the fusion of natural language processing and computer vision, these models may provide insights into patient-specific surgical planning, possibly enhancing precision, safety, and patient outcomes.³⁷

Assessing the LLMs' recommendations for managing various technical aspects of rhinoplasty proved interesting. Multiple factors influence a plastic surgeon's approach to nasal reconstruction, wherein applying inappropriate surgical methods can adversely affect patient outcomes. Consequently, AI tools that provide erroneous recommendations may have legal and ethical ramifications that could implicate both the treating team and the software developer.³⁸ Considering the limitations uncovered in this study, all 3 LLMs necessitate significant enhancements before their practical application in rhinoplasty management. Therefore, users should consult experienced plastic surgeons in addition to the 3 models' recommendations.

In summary, the LLMs demonstrated a moderate grasp of the rhinoplasty-related questions posed, providing logical and easily understandable answers. As LLMs, their primary function revolves around predicting the likelihood of a word sequence based on the context provided by preceding words. While this ability has allowed the LLMs to achieve impressive feats, it also limits their capacity to offer in-depth information. Consequently, this study underscores the limitations of LLMs in delivering detailed knowledge on specialized surgical subjects. Moving forward, the creation of a "scholar"-type LLM tailored for physicians and similar professionals would be valuable. Google Scholar, PubMed, and other databases curate peer-reviewed academic literature, which is highly useful but not directly actionable in a clinical setting. A specialized LLM could potentially fill this gap by distilling complex, peer-reviewed medical information into actionable information for health-care providers. We anticipate that many of the current limitations will be resolved in time given the rapidly changing nature of this technology, yet, at this point of time, further

refinement is necessary for these models to be effectively integrated into clinical practice.

CONCLUSIONS

Overall, this study demonstrates that using LLMs such as ChatGPT, BARD, and Bing AI for acquiring detailed information on specialized surgical procedures like rhinoplasty has inherent limitations. While these models can offer pertinent and accessible information, inconsistencies and superficial content may be present. It is imperative to scrupulously evaluate the information provided by these AI systems and corroborate it with evidence-based sources and expert insights to ensure accuracy and reliability in rhinoplasty presurgical planning, decision making, and patient education.

Supplemental Material

This article contains [supplemental material](http://www.asjopenforum.com) located online at www.asjopenforum.com.

Acknowledgments

Dr Seth and Mr Lim have contributed equally for first authorship.

Disclosures

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Funding

The authors received no financial support for the research, authorship, and publication of this article, including payment of the article processing charge.

REFERENCES

1. Lee MK, Most SP. Evidence-based medicine: rhinoplasty. *Facial Plast Surg Clin*. 2015;23(3):303-312. doi: [10.1016/j.fsc.2015.04.004](https://doi.org/10.1016/j.fsc.2015.04.004)
2. Fichman M, Piedra Buena IT. Rhinoplasty. In: *StatPearls*. StatPearls Publishing; 2023:1-3.
3. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research using ChatGPT. *Aesthet Surg J*. 2023;43(8):930-937. doi: [10.1093/asj/sjad069](https://doi.org/10.1093/asj/sjad069)
4. Seth I, Cox A, Xie Y, et al. Evaluating Chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J*. 2023;43(10):1126-1135. doi: [10.1093/asj/sjad140](https://doi.org/10.1093/asj/sjad140)
5. Seth I, Xie Y, Rodwell A, et al. Exploring the role of a large language model on carpal tunnel syndrome management: an observation study of ChatGPT. *J Hand Surg Am*. 2023. doi: [10.1016/j.jhssa.2023.07.003](https://doi.org/10.1016/j.jhssa.2023.07.003) [Epub ahead of print].

6. Seth I, Lim B, Xie Y, Hunter-Smith DJ, Rozen WM. Exploring the role of artificial intelligence ChatBot on the management of scaphoid fractures. *J Hand Surg Eur Vol.* 2023;48(8):814-818. doi: [10.1177/17531934231169858](https://doi.org/10.1177/17531934231169858)
7. Seth I, Kenney PS, Bulloch G, Hunter-Smith DJ, Thomsen JB, Rozen WM. Artificial or augmented authorship? A conversation with a Chatbot on base of thumb arthritis. *Plast Reconstr Surg Global Open.* 2023;11(5):e4999. doi: [10.1097/GOX.0000000000004999](https://doi.org/10.1097/GOX.0000000000004999)
8. Bhattacharya K, Bhattacharya AS, Bhattacharya N, Yagnik VD, Garg P, Kumar S. ChatGPT in surgical practice—a new kid on the block. *Indian J Surg.* 2023;1-4. doi: [10.1007/s12262-023-03727-x](https://doi.org/10.1007/s12262-023-03727-x)
9. Gupta R, Herzog I, Weisberger J, Chao J, Chaiyasate K, Lee ES. Utilization of ChatGPT for plastic surgery research: friend or foe? *J Plast Reconstr Aesthet Surg.* 2023;80:145-147. doi: [10.1016/j.bjps.2023.03.004](https://doi.org/10.1016/j.bjps.2023.03.004)
10. Cox A, Seth I, Xie Y, Hunter-Smith DJ, Rozen WM. Utilizing ChatGPT-4 for providing medical information on blepharoplasties to patients. *Aesthet Surg J.* 2023;43(8):NP658-NP662. doi: [10.1093/asj/sjad096](https://doi.org/10.1093/asj/sjad096)
11. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthet Plast Surg.* 2023. doi: [10.1007/s00266-023-03338-7](https://doi.org/10.1007/s00266-023-03338-7) [Epub ahead of print].
12. Xie Y, Seth I, Rozen WM, Hunter-Smith DJ. Evaluation of the artificial intelligence ChatBot on breast reconstruction and its efficacy in surgical research: a case study. *Aesthetic Plast Surg.* 2023. doi: [10.1007/s00266-023-03443-7](https://doi.org/10.1007/s00266-023-03443-7) [Epub ahead of print].
13. Toriumi DM, Josen J, Weinberger M, Tardy ME Jr. Use of alar batten grafts for correction of nasal valve collapse. *Arch Otolaryngol Head Neck Surg.* 1997;123(8):802-808. doi: [10.1001/archotol.1997.01900080034002](https://doi.org/10.1001/archotol.1997.01900080034002)
14. Most SP. Anterior septal reconstruction: outcomes after a modified extracorporeal septoplasty technique. *Arch Facial Plast Surg.* 2006;8(3):202-207. doi: [10.1001/archfaci.8.3.202](https://doi.org/10.1001/archfaci.8.3.202)
15. André RF, Vuyk HD. The “butterfly graft” as a treatment for internal nasal valve incompetence. *Plast Reconstr Surg.* 2008;122(2):73e-74e. doi: [10.1097/PRS.0b013e31817d60cc](https://doi.org/10.1097/PRS.0b013e31817d60cc)
16. Teymoortash A, Fasanla JA, Sazgar AA. The value of spreader grafts in rhinoplasty: a critical review. *Eur Arch Otorhinolaryngol.* 2012;269(5):1411-1416. doi: [10.1007/s00405-011-1837-y](https://doi.org/10.1007/s00405-011-1837-y)
17. Spielmann PM, White PS, Hussain S. Surgical techniques for the treatment of nasal valve collapse: a systematic review. *Laryngoscope.* 2009;119(7):1281-1290. doi: [10.1002/lary.20495](https://doi.org/10.1002/lary.20495)
18. Kridel RW, Scott BA, Foda HM. The tongue-in-groove technique in septorhinoplasty. A 10-year experience. *Arch Facial Plast Surg.* 1999;1(4):246-256; discussion 257-258. doi: [10.1001/archfaci.1.4.246](https://doi.org/10.1001/archfaci.1.4.246)
19. Toriumi DM. New concepts in nasal tip contouring. *Arch Facial Plast Surg.* 2006;8(3):156-185. doi: [10.1001/archfaci.8.3.156](https://doi.org/10.1001/archfaci.8.3.156)
20. Passàli D, Passàli FM, Damiani V, Passàli GC, Bellussi L. Treatment of inferior turbinate hypertrophy: a randomized clinical trial. *Ann Otol Rhinol Laryngol.* 2003;112(8):683-688. doi: [10.1177/000348940311200806](https://doi.org/10.1177/000348940311200806)
21. Sinno S, Mehta K, Lee Z-H, Kidwai S, Saadeh PB, Lee MR. Inferior turbinate hypertrophy in rhinoplasty: systematic review of surgical techniques. *Plast Reconstr Surg.* 2016;138(3):419e-429e. doi: [10.1097/PRS.0000000000002433](https://doi.org/10.1097/PRS.0000000000002433)
22. Zhang K, Pipaliya RM, Miglani A, Nguyen SA, Schlosser RJ. Systematic review of surgical interventions for inferior turbinate hypertrophy. *Am J Rhinol Allergy.* 2023;37(1):110-122. doi: [10.1177/19458924221134555](https://doi.org/10.1177/19458924221134555)
23. Park SS. Fundamental principles in aesthetic rhinoplasty. *Clin Exp Otorhinolaryngol.* 2011;4(2):55-66. doi: [10.3342/ceo.2011.4.2.55](https://doi.org/10.3342/ceo.2011.4.2.55)
24. Şirinoğlu H. The effect of the short and floating columellar strut graft and septocolumellar suture on nasal tip projection and rotation in primary open approach rhinoplasty. *Aesthetic Plast Surg.* 2017;41(1):146-152. doi: [10.1007/s00266-016-0727-5](https://doi.org/10.1007/s00266-016-0727-5)
25. Hontanilla B, Cabello A, Olivas J. A predictable approach for osteotomy in rhinoplasty: a new concept of open external osteotomy. *Plast Reconstr Surg Glob Open.* 2016;4(6):e764. doi: [10.1097/GOX.0000000000000738](https://doi.org/10.1097/GOX.0000000000000738)
26. Allauzen C, Mohri M, Roark B. Generalized algorithms for constructing statistical language models. 2003:40-47.
27. Chinski H, Lerch R, Tournour D, Chinski L, Caruso D. An artificial intelligence tool for image simulation in rhinoplasty. *Facial Plast Surg.* 2021;38(2):201-206. doi: [10.1055/s-0041-1729911](https://doi.org/10.1055/s-0041-1729911)
28. Gupta R, Pande P, Herzog I, et al. Application of ChatGPT in cosmetic plastic surgery: ally or antagonist? *Aesthet Surg J.* 2023;43(7):NP587-NP590. doi: [10.1093/asj/sjad042](https://doi.org/10.1093/asj/sjad042)
29. Saleem H, Li Y, Ali Z, Ayyoub M, Wang Y, Mehreen A. Big data use and its outcomes in supply chain context: the roles of information sharing and technological innovation. *J Enterp Inf Manage.* 2021;34(4):1121-1143. doi: [10.1108/JEIM-03-2020-0119](https://doi.org/10.1108/JEIM-03-2020-0119)
30. Salvagno M, Taccone FS, Gerli AG. Correction to: can artificial intelligence help for scientific writing? *Crit Care.* 2023;27(1):99. doi: [10.1186/s13054-023-04390-0](https://doi.org/10.1186/s13054-023-04390-0)
31. Eldaly AS, Avila FR, Torres-Guzman RA, et al. Simulation and artificial intelligence in rhinoplasty: a systematic review. *Aesthetic Plast Surg.* 2022;46(5):2368-2377. doi: [10.1007/s00266-022-02883-x](https://doi.org/10.1007/s00266-022-02883-x)
32. Li J, Dada A, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. medRxiv. 2023. doi: [10.1101/2023.03.30.23287899](https://doi.org/10.1101/2023.03.30.23287899)
33. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med.* 2023;21(1):269. doi: [10.1186/s12967-023-04123-5](https://doi.org/10.1186/s12967-023-04123-5)
34. Baumann I. Quality of life before and after septoplasty and rhinoplasty. *GMS Curr Top Otorhinolaryngol Head Neck Surg.* 2010;9:Doc06. doi: [10.3205/cto000070](https://doi.org/10.3205/cto000070)

35. Cole RP, Shakespeare V, Shakespeare PG, Hobby JA. Measuring outcome in low-priority plastic surgery patients using quality of life indices. *Br J Plast Surg.* 1994;47(2): 117-121. doi: [10.1016/0007-1226\(94\)90170-8](https://doi.org/10.1016/0007-1226(94)90170-8)
36. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018. doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)
37. Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. 2019. doi: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692)
38. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J Assoc Inf Sci Technol.* 2023;74(5):570-581. doi: [10.1002/asi.24750](https://doi.org/10.1002/asi.24750)