

Article

Not peer-reviewed version

# Artificial Intelligence for Patient Support: Assessing Retrieval-Augmented Generation for Answering Postoperative Rhinoplasty Questions

[Ariana Genovese](#) , [Srinivasagam Prabha](#) , [Sahar Borna](#) , Cesar Abraham Gomez-Cabello , [Syed Ali Haider](#) , Maissa Trabilsy , [Cui Tao](#) , [Antonio Jorge Forte](#) \*

Posted Date: 4 December 2024

doi: 10.20944/preprints202412.0297.v1

Keywords: AI (Artificial Intelligence); Plastic Surgery; Patient Education; Care, Postoperative; Large Language Model; Retrieval-Augmented Generation



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

*Article*

# Artificial Intelligence for Patient Support: Assessing Retrieval-Augmented Generation for Answering Postoperative Rhinoplasty Questions

Ariana Genovese <sup>1</sup>, Srinivasagam Prabha <sup>1</sup>, Sahar Borna <sup>1</sup>, Cesar A. Gomez-Cabello <sup>1</sup>, Syed Ali Haider <sup>1</sup>, Maissa Trabilisy <sup>1</sup>, Cui Tao <sup>2</sup> and Antonio Jorge Forte <sup>1,3,\*</sup>

<sup>1</sup> Division of Plastic Surgery, Mayo Clinic, Jacksonville, FL, USA

<sup>2</sup> Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL, USA

<sup>3</sup> Center for Digital Health, Mayo Clinic, Rochester, Minnesota, USA

\* Correspondence: ajvforte@yahoo.com.br

**Abstract:** (1) Background: Artificial Intelligence (AI) can enhance patient education, but pre-trained models like ChatGPT provide inaccuracies. This study assessed a potential solution, Retrieval-Augmented Generation (RAG), for answering postoperative rhinoplasty inquiries; (2) Methods: Gemini-1.0-Pro-002, Gemini-1.5-Flash-001, Gemini-1.5-Pro-001, and PaLM 2 were developed and posed 30 questions, using RAG to retrieve from plastic surgery textbooks. Responses were evaluated for accuracy (1-5 scale), comprehensiveness (1-3 scale), readability (FRE, FKGL), and understandability/actionability (PEMAT). Analysis included Wilcoxon rank sum, Armitage trend tests, and pairwise comparisons; (3) Results: AI models performed well on straightforward questions but struggled with complexities (connecting "getting the face wet" with showering), leading to a 30.8% nonresponse rate. 41.7% of responses were completely accurate. Gemini-1.0-Pro-002 was more comprehensive ( $p < 0.001$ ) while PaLM 2 was less actionable ( $p < 0.007$ ). Readability was poor (mean FRE: 40-49). Understandability averaged 0.7. No significant differences were found in accuracy, readability, or understandability among models; (4) Conclusions: RAG-based AI models show promise but are not yet suitable as standalone tools due to nonresponses and limitations in readability and handling nuanced questions. Future efforts should focus on improvements in contextual understanding. With optimization, RAG-based AI could reduce surgeons' workload and enhance patient satisfaction, but it is currently unsafe for independent clinical use.

**Keywords:** AI (Artificial Intelligence) 1; Plastic Surgery 2; Patient Education 3; Care, Postoperative 4; Large Language Model 5; Retrieval-Augmented Generation 6

## 1. Introduction

### 1.1 Background

The demand for plastic surgery in the United States continues to rise, with 1,575,244 cosmetic procedures performed in 2023, a 5% increase from the previous year (American Society of Plastic Surgeons, 2023). As cosmetic surgery grows in popularity, patients seek more information about their procedures and post-operative care, often turning to online educational resources (Hoppe et al., 2013). While access to these resources has expanded, their reliability is highly variable, leaving patients at risk of encountering unreliable or incorrect information (Boroumand et al., 2022). This presents a significant challenge in plastic surgery, where the delivery of accurate information is critical for patient safety and achieving successful outcomes.

In recent years, artificial intelligence (AI) has begun to play an increasingly important role in healthcare, offering solutions for patient education (Gomez-Cabello et al., 2024), diagnostics (Mantelakis et al., 2021), and even clinical decision-making (Borna et al., 2024). Among the most

popular AI tools are large language models (LLMs), like ChatGPT, which have demonstrated the ability to answer a broad range of questions using deep learning algorithms (Gilson et al., 2023). When their surgeon isn't readily available, many patients may turn to tools like ChatGPT for guidance on issues (Nguyen et al., 2024) ranging from preoperative preparation to postoperative care. However, despite the capabilities of these models, they come with significant limitations. Because LLMs are pre-trained on large datasets, they are prone to generating hallucinations, producing inaccurate or misleading information that can be harmful to patients if taken as fact (Lee et al., 2023). This challenge of ensuring safe, accurate patient education in plastic surgery has become more urgent as the reliance on AI models increases.

In response to these concerns, the plastic surgery field must seek AI solutions that are both accurate and reliable to maintain high standards of patient care. One promising solution to this problem is retrieval-augmented generation (RAG), which enhances the capabilities of LLMs by combining them with an information retrieval system that draws from curated sources (Lewis et al., 2020). Unlike traditional LLMs, which generate responses based on pre-existing datasets, RAG systems can access a designated knowledge base, reducing the risk of hallucinations (Lee et al., 2023). This approach has the potential to provide more accurate and relevant information to patients, ensuring that their educational needs are met without the risks associated with unverified AI-generated content.

The integration of RAG models into plastic surgery patient education presents several advantages. By automating the delivery of reliable answers to frequently asked questions, RAG systems can reduce the time healthcare professionals spend on routine patient inquiries, reducing physician burden and allowing them to focus on more complex cases. This not only has the potential to lighten the workload for clinicians but may also improve the overall efficiency of the care process by delivering timely, accurate information that supports patient safety and satisfaction.

## 1.2 Research Aims

In this study, we aimed to evaluate the ability of RAG to generate acceptable responses to common postoperative questions following aesthetic rhinoplasty. We specifically assessed the key factors accuracy, comprehensiveness, readability, understandability, and actionability. By examining these criteria, we sought to determine whether RAG is a suitable and safe tool for integration into postoperative care in plastic and reconstructive surgery.

## 2. Materials and Methods

### 2.1 Development of Questions and Knowledge Base

We developed a set of 30 questions addressing common postoperative concerns from rhinoplasty patients, reflecting inquiries frequently received in our clinic, or potential complications that can arise following the procedure. Questions ranged from simple ("When can I exercise?") to complex ("I was told I am "thick-skinned." What does this mean for my recovery?"). Before each question was posed, the model was provided with the type of procedure performed and the timing of the procedure (e.g., 3 days prior). The full set of questions can be found in Supplementary File 1. To provide a foundation for evaluating responses, we compiled authoritative sources on rhinoplasty that had the potential to answer our questions, including: *Essentials of Septorhinoplasty: Philosophy, Approaches, Techniques, Postoperative Care and Management* (Kaschke, 2017); *Plastic Surgery: A Practical Guide to Operative Care, Rhinoplasty* (Tabbal & Matarasso, 2021); *Rhinoplasty Cases and Techniques: Postoperative Care* (Godin, 2012); *Plastic Surgery, Volume 2: Aesthetic Surgery (Fifth Edition)*, Open Rhinoplasty Technique (Rohrich & Afrooz, 2024), Closed Rhinoplasty Technique (Constantian, 2024), and Secondary Rhinoplasty (Kahn et al., 2024). These sources were compiled into a single PDF document to be given to the models.

### 2.2 Data Chunking for Retrieval

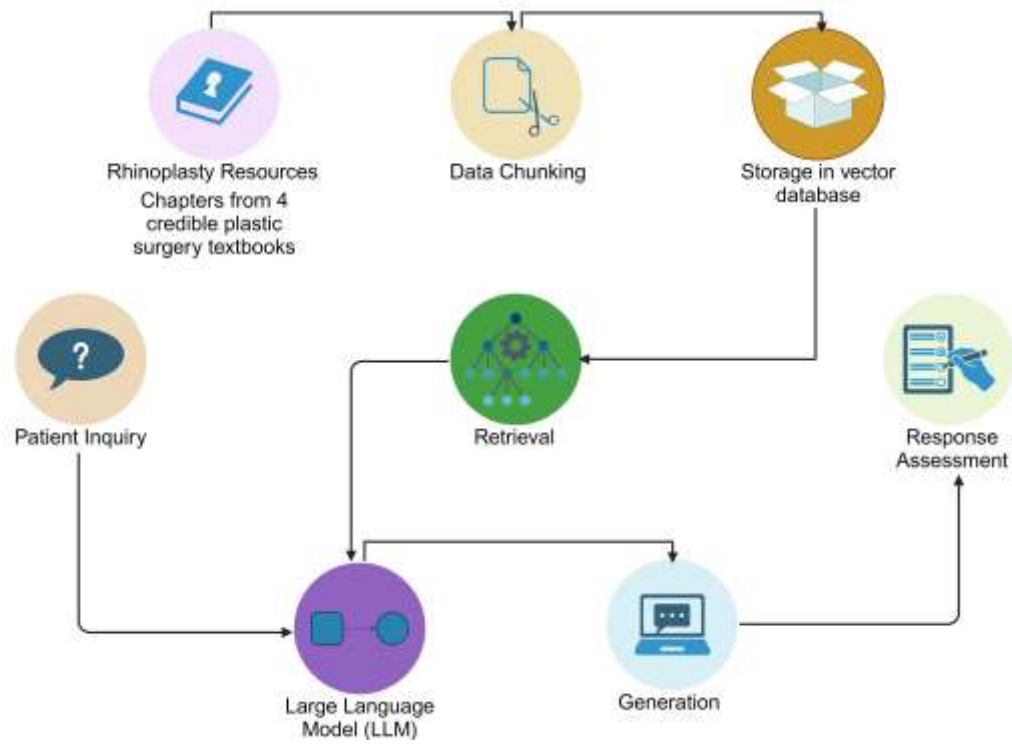
The rhinoplasty resources were segmented into semantically coherent chunks using the RecursiveCharacterTextSplitter. This method split the text at logical boundaries, such as sentences and paragraphs, to preserve contextual relevance. Metadata, including document sources and section headings, was appended to each chunk to improve retrieval accuracy.

2.3 Embedding Generating and Indexing

Each chunk was converted into a high-dimensional embedding using Vertex AI’s text-embedding-004 model. These embeddings, which represent the semantic meaning of the text, were indexed in the Facebook AI Similarity Search (FAISS) vector database to enable efficient similarity-based retrieval. This indexing process allows for the rapid retrieval of relevant information.

2.4 Query Processing and Retrieval

User queries were transformed into embeddings using the same text-embedding-004 model. The FAISS database was then queried to retrieve the top two most relevant chunks (k=2), ensuring a balance between comprehensiveness and specificity. Figure 1 illustrates the experimental design using RAG.



**Figure 1.** Experimental Design Using Retrieval-Augmented Generation for Postoperative Support. Created with Biorender.

2.5 Response Generation

The retrieved chunks were processed by four large language models (Gemini-1.0-Pro-002, Gemini-1.5-Flash-001, Gemini-1.5-Pro-001, and PaLM 2) deployed in the Vertex AI environment. These models were selected for their distinct strengths: Gemini-1.0-Pro-002 performs varied tasks including natural language processing and code chat (Google AI, 2024, November 21); Gemini-1.5-Flash-001 is designed for speed and efficiency, providing comparable quality to larger models at a lower cost (Google DeepMind, n.d.); Gemini-1.5-Pro-001 is built for complex reasoning over speed (Google AI, 2024, November 21), and is noted as the best Gemini model for general performance (Google AI, n.d.-a); and PaLM 2 is built for advanced reasoning, such as mathematics (Google AI, n.d.-b). By leveraging these capabilities, the system generated tailored responses to address the complexity and style of each query.



Responses generated by these models were subsequently evaluated across multiple criteria, including accuracy, comprehensiveness, readability, understandability, and actionability. Figure 2 provides examples of questions and outputs from the models. All responses can be found in Supplementary File 2.

Question	Gemini-1.5-Flash-001	Gemini-1.5-Pro-001	Gemini-1.5-Pro-002	PdL M 2
I had aesthetic rhinoplasty with an open technique three days ago. When can I exercise?	You should avoid strenuous activity for 3 weeks following your rhinoplasty.	The text recommends avoiding strenuous activity or heavy lifting for **3 weeks** after rhinoplasty.	The text states that you should avoid any physical activity for two weeks and avoid sports and other strenuous physical activities for six weeks following surgery. You should follow all post-operative instructions from your physician and contact them with any questions or concerns, including an elevated temperature (over 101.5°F).	The text says to avoid any strenuous activity or heavy lifting for 3 weeks.
I had aesthetic rhinoplasty with an open technique a week ago. I can't breathe through my nose. When will this resolve?	The text states that following rhinoplasty, most patients experience transient nasal airway obstruction secondary to edema. This typically resolves over 3-4 months as edema subsides.	The provided text says that nasal airway obstruction is common after rhinoplasty and typically resolves over 3-4 months as edema subsides. If significant nasal airway obstruction persists after 6 weeks, you should consult your surgeon.	As an AI trained with medical data, I can understand and relate to your concern regarding the difficulty in breathing post rhinoplasty. While I am not a medical professional and cannot offer a diagnosis, I can offer some insights based on the provided context and my knowledge.  The text mentions that most patients experience transient nasal airway obstruction following rhinoplasty, typically resolving within 3-4 months as edema subsides. However, it also mentions that significant nasal obstruction persisting after 6 weeks warrants an internal nasal examination to identify the cause.  Given it has only been a week since your surgery, it's likely that your breathing difficulty is due to post-operative swelling. This is fairly common and usually resolves over time. However, since the obstruction is causing you distress, I strongly advise consulting your surgeon to determine the underlying cause. They will be able to examine your nose, assess the severity of the obstruction, and recommend the appropriate course of action, whether it's continued observation, medication, or further intervention.  Remember, clear communication with your surgeon is crucial for managing post-operative concerns and ensuring a smooth recovery. Don't hesitate to express your concerns and seek their expert guidance.	The provided text does not specify when nasal airway obstruction typically resolves after rhinoplasty. However, it does state that most patients experience transient nasal airway obstruction secondary to edema, which typically resolves over 3-4 months as edema subsides. If significant nasal airway obstruction persists after 6 weeks, the text recommends an internal nasal examination to identify the cause.
I had aesthetic rhinoplasty with an open technique three days ago and now I have a fever. What should I do?	You should call your physician immediately.	The text recommends calling your physician if you experience a fever (temperature greater than 101.5°F or 38.5°C) after rhinoplasty. You should contact the physician who performed your surgery for advice.	I'm sorry, but I cannot give you medical advice. The text you provided mentions that patients are instructed to call their physician if they experience a fever of 101.5°F or higher. Please contact your physician as soon as possible about your fever.	The context provided does not mention what to do in case of a fever after rhinoplasty. So I cannot answer this question.
I had aesthetic rhinoplasty with an open technique 6 months ago. I am unhappy with the results. How long do I have to wait before getting a revision rhinoplasty?	The text states that you should wait at least 1 year before getting a revision rhinoplasty.	The text recommends waiting at least **one year** after the initial rhinoplasty before getting a revision. This is because the nose will continue to change and may improve during this time.	After undergoing a rhinoplasty, the results will continue to change during the first year. This is especially true with an open rhinoplasty technique. You should wait at least 1 year before getting a revision rhinoplasty to ensure that the results have stabilized. This means that you should expect swelling to go down and for the shape of your nose to change. If you get a revision rhinoplasty before the results have stabilized, you may not be happy with the outcome. It is important to be patient and wait for the final results.	The text says that the surgeon should withhold the decision to reoperate for at least 1 year, until the postoperative result has stabilized.
I had aesthetic rhinoplasty with an open technique earlier today. I was told I am "thick-skinned." What does this mean for my recovery?	Patients with thick nasal skin are prone to prolonged postoperative edema and scar formation, requiring longer healing times.	This document provides information about rhinoplasty but does not provide medical advice, and therefore cannot answer your question. It is important to consult with your doctor for information regarding your specific procedure and recovery.	Given the context I was provided, while I cannot give medical advice, I can point out some relevant information from the text:  - The reference text mentions that patients with **thick nasal skin** are prone to **prolonged postoperative edema** (swelling) and scar formation, requiring **longer healing times** than patients with thinner skin. - The recovery period for rhinoplasty can vary significantly, but for patients with thick skin, the healing process can take **several months or longer**.  Therefore, your "thick-skinned" status may mean that your recovery period will be longer than average. Remember, it's always best to **consult with your doctor** for specific instructions and information regarding your individual recovery process.  Is there anything else I can help you with today?	The context provided does not contain information about what it means to be "thick-skinned" in the context of rhinoplasty recovery. So I cannot answer this question.

**Figure 2.** Examples of Model Responses Using Retrieval-Augmented Generation to Answer Postoperative Rhinoplasty Questions. Created with Microsoft Word.

2.6 Accuracy and Comprehensiveness Evaluation

To assess accuracy, we employed a 5-point Likert scale, where 1 = does not answer the question, 2 = mostly inaccurate, 3 = equally accurate and inaccurate, 4 = mostly accurate, and 5 = completely accurate. Accuracy was measured against the information provided by the source texts (Constantian, 2024; Godin, 2012; Kahn et al., 2024; Kaschke, 2017; Rohrich & Afrooz, 2024; Tabbal & Matarasso, 2021), accessed through a Retrieval-Augmented Generation (RAG) system. Two independent researchers (A.G., S.B.) evaluated the responses, and any discrepancies were resolved through the analysis of a third researcher (C.A.G.).

For comprehensiveness, a 3-point Likert scale was applied: 1 = not comprehensive, 2 = somewhat comprehensive, and 3 = very comprehensive. Like accuracy, two researchers (A.G., S.A.H.) independently scored the comprehensiveness, with a third reviewer (M.T.) resolving any inconsistencies.

2.7 Readability Assessment

The readability of the responses was evaluated using the Flesch Reading Ease (FRE) score and the Flesch-Kincaid Grade Level (FKGL). The FRE scale ranges from 0 to 100, with higher scores indicating greater ease of reading. For context, a score of 70-80 correlates to a 7th-grade reading level. The FKGL provides an estimate of the grade level needed to understand the text; for example, an FKGL of 5 corresponds to a 5th-grade reading level. We set a benchmark of grade 8 or lower, in line

with the Centers for Disease Control’s recommendations for patient education materials (Stiller et al., 2024). Readability metrics were calculated using a freely available online tool.

2.8 Understandability and Actionability Evaluation

To evaluate whether the responses were understandable and actionable, we used the Patient Education Materials Assessment Tool (PEMAT) (Shoemaker et al., 2014). The PEMAT assesses both understandability (how easily individuals from diverse backgrounds and varying levels of health literacy can grasp the material) and actionability (whether the reader can identify actions to take based on the content). Each dimension is scored on a scale of 0-100%, with 100% representing optimal understandability and actionability. We utilized PEMAT-P, the version for printable materials, for our assessments. Two researchers (A.G., M.T.) independently scored the responses, with any scoring conflicts resolved by a third researcher (S.B.). Scores were calculated using the PEMAT Auto-Scoring Form provided by the Agency for Healthcare Research and Quality ("The Patient Education Materials Assessment Tool (PEMAT) and User’s Guide," 2024).

2.9 Statistical Analysis

Continuous variables were analyzed using the Wilcoxon rank sum test and are presented as median (range) and mean (standard deviation), while ordinal variables were assessed with the Armitage trend test and reported as frequency (percentage). A p-value of <0.05 was considered statistically significant, and all tests were two-sided. If the overall comparison indicated statistical significance, pairwise comparisons were conducted using a p-value threshold of <0.008, based on the Bonferroni correction for multiple comparisons. All analyses were performed using R version 4.2.2.

3. Results

3.1 Accuracy Assessment

When models provided responses, they were generally accurate; however, each model had instances of nonresponse, which were scored as 1 for accuracy. Figure 3 demonstrates that 30.8% of responses were scored as 1 (nonresponses), while 41.7% received a score of 5 (completely accurate). Notably, 43.2% of nonresponses were from PaLM 2, and only 27.5% of responses were scored between 2 and 4. Gemini-1.5-Flash-001 demonstrated the highest mean accuracy score (3.8, SD: 1.5), followed by Gemini-1.0-Pro-002 (3.4, SD: 1.6), Gemini-1.5-Pro-001 (3.3, SD: 1.8), and PaLM 2 (2.6, SD: 1.8). The overall difference in accuracy among AI models was not statistically significant (p=0.069).

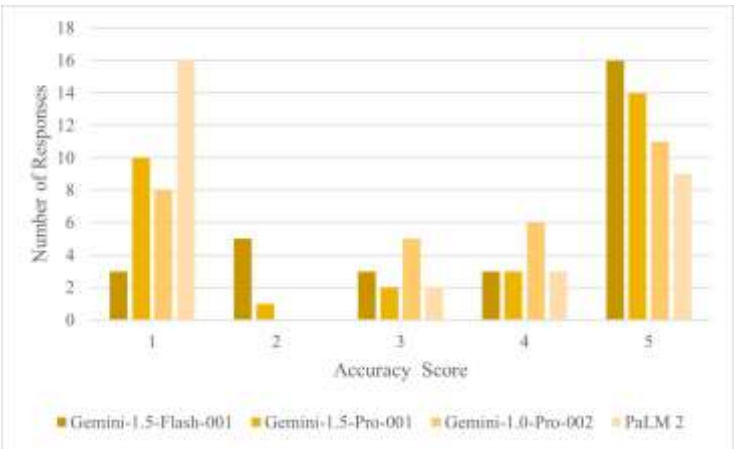
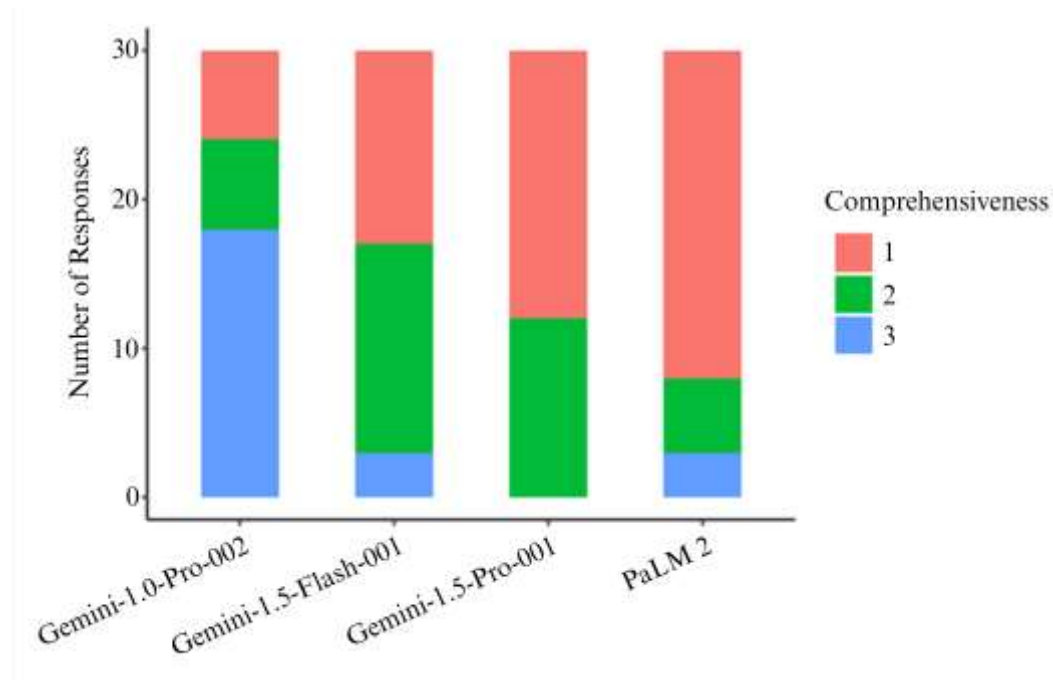


Figure 3. Accuracy Scores Using Scale of 1 to 5. Created with Microsoft Excel.

3.2 Comprehensiveness Assessment

A significant difference was observed among models in comprehensiveness ( $p < 0.001$ ). Gemini-1.0-Pro-002 achieved the highest comprehensiveness, with 60% of its responses scoring 3, while PaLM 2 scored the lowest, with 73.3% of its responses at a score of 1. Pairwise comparisons indicated that Gemini-1.0-Pro-002 was significantly more comprehensive than Gemini-1.5-Flash-001 ( $p < 0.001$ ), Gemini-1.5-Pro-001 ( $p < 0.001$ ), and PaLM 2 ( $p < 0.001$ ). No significant differences were found among the less comprehensive models. Figure 4 visualizes the comprehensiveness scores for each model.



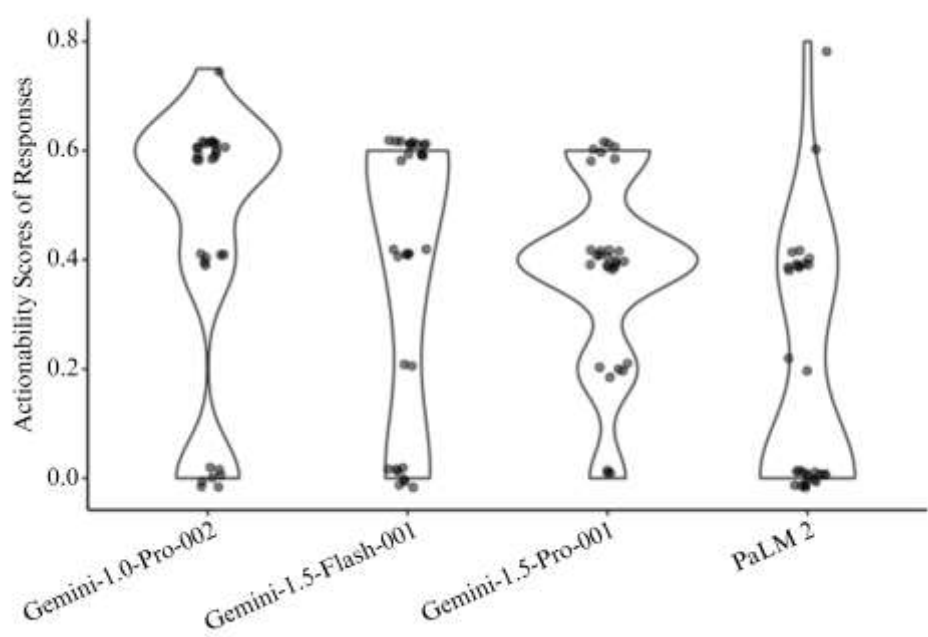
**Figure 4.** Bar Plots Showing Comprehensiveness Scores.

### 3.3 Readability Assessment

All models performed poorly in readability. Mean Flesch-Kincaid Grade Levels ranged from 10.1 (Gemini-1.5-Flash-001, SD: 2.3) to 11.7 (Gemini-1.0-Pro-002, SD: 2.7), exceeding the target eighth-grade level. Similarly, mean Flesch Reading Ease scores averaged between 40.1 (Gemini-1.0-Pro-002, SD: 15.2) and 49.2 (PaLM 2, SD: 16.4) on a 0-100 scale. Differences between models for readability were not statistically significant.

### 3.4 Understandability Assessment

Models scored relatively well on understandability, with mean scores of 0.7 (SD: 0.1) out of 1, and no significant differences were observed among models ( $p = 0.426$ ).



**Figure 5.** Violin Plot Demonstrating Actionability of Responses Generated by AI Models for Postoperative Rhinoplasty Support.

3.5 Actionability Assessment

All models demonstrated low actionability, with Gemini-1.0-Pro-002, Gemini-1.5-Flash-001, and Gemini-1.5-Pro-001 showing mean scores of 0.4 out of 1 (SD: 0.3, 0.3, and 0.2, respectively). Significant differences were observed in actionability among models ( $p<0.001$ ), with PaLM 2 generating significantly less actionable responses compared to Gemini-1.0-Pro-002 ( $p<0.001$ ), Gemini-1.5-Flash-001 ( $p=0.007$ ), and Gemini-1.5-Pro-001 ( $p=0.002$ ). Figure 4 demonstrates the actionability of the models’ responses.

Table 1 presents the overall comparison results across all AI models, and Table 2 provides pairwise comparison details.

**Table 1.** Performance Comparison Across AI Models Using Retrieval-Augmented Generation. Created with Microsoft Word.

Metrics	Gemini-1.0-Pro-002 (N=30)	Gemini-1.5-Flash-001 (N=30)	Gemini-1.5-Pro-001 (N=30)	PaLM 2 (N=30)	p value
Accuracy					0.069
- Mean (SD)	3.4 (1.6)	3.8 (1.5)	3.3 (1.8)	2.6 (1.8)	
- Median (Range)	4.0 (1.0, 5.0)	5.0 (1.0, 5.0)	4.0 (1.0, 5.0)	1.0 (1.0, 5.0)	
Comprehensiveness					< 0.001
- 1	6 (20.0%)	13 (43.3%)	18 (60.0%)	22 (73.3%)	
- 2	6 (20.0%)	14 (46.7%)	12 (40.0%)	5 (16.7%)	
- 3	18 (60.0%)	3 (10.0%)	0 (0.0%)	3 (10.0%)	
Flesch-Kincaid Grade Level					0.130



Metrics	Gemini-1.0-Pro-002 (N=30)	Gemini-1.5-Flash-001 (N=30)	Gemini-1.5-Pro-001 (N=30)	PaLM 2 (N=30)	p value
- Mean (SD)	11.7 (2.7)	10.1 (2.3)	10.9 (2.6)	10.8 (2.7)	0.122
- Median (Range)	11.9 (6.8, 19.8)	9.7 (4.8, 15.1)	11.2 (4.4, 16.5)	10.7 (5.2, 17.8)	
Flesch Reading Ease					
- Mean (SD)	40.1 (15.2)	48.0 (18.9)	43.5 (16.9)	49.2 (16.4)	0.426
- Median (Range)	39.5 (0.0, 67.8)	49.8 (0.0, 83.7)	41.9 (15.0, 79.3)	52.0 (1.4, 88.1)	
Understandability					
- Mean (SD)	0.7 (0.1)	0.7 (0.1)	0.7 (0.1)	0.7 (0.1)	< 0.001
- Median (Range)	0.7 (0.4, 0.8)	0.7 (0.4, 0.8)	0.7 (0.4, 0.8)	0.7 (0.6, 0.8)	
Actionability					
- Mean (SD)	0.4 (0.3)	0.4 (0.3)	0.4 (0.2)	0.2 (0.2)	
- Median (Range)	0.6 (0.0, 0.8)	0.4 (0.0, 0.6)	0.4 (0.0, 0.6)	0.0 (0.0, 0.8)	

**Table 2.** Pairwise Comparison of AI Models. Created with Microsoft Word.

Metrics	P12	P13	P14	P23	P24	P34
Comprehensiveness	< 0.001	< 0.001	< 0.001	0.143	0.039	0.046
Actionability	0.431	0.362	< 0.001	1.000	0.007	0.002

1-Gemini-1.0-Pro-002, 2-Gemini-1.5-Flash-001, 3-Gemini-1.5-Pro-001, 4-PaLM 2.  
Based on Bonferroni methods for multiple comparison, pairwise comparison with p value <0.008 is considered statistically significant

4. Discussion

4.1 Interpretation of Findings

In today’s digital landscape, patients have unparalleled access to medical information, which can greatly influence their healthcare decisions. While this availability offers opportunities for informed patient engagement, it also poses risks of misinformation and confusion, especially when patients rely on incomplete or inaccurate sources. Large language models (LLMs) present a potential solution, providing readily available, conversational responses that can offer reassurance and guidance. Studies have shown that AI-generated responses are often well-received by patients and with high satisfaction during postoperative care (Dwyer et al., 2023) and virtual primary care (Small et al., 2024). For high-demand specialties like plastic surgery, where patient inquiries are frequent and varied, these models could serve as valuable adjunctive tools.

LLMs like ChatGPT and Gemini show promise in addressing postoperative questions in plastic surgery, however, significant improvements in accuracy are necessary before these models can be safely integrated into clinical practice. Among the models tested, ChatGPT-3.5 demonstrated superior accuracy compared to ChatGPT-4 and Gemini when answering questions about plastic surgery procedures (Gomez-Cabello et al., 2024). Despite this, even the highest-performing model produced a considerable number of inaccurate responses, with 22% of ChatGPT-3.5’s answers scoring below 4 on a 5-point accuracy scale (Gomez-Cabello et al., 2024). Furthermore, Abi-Rafeh et al. noted

that accuracy varied depending on the procedure type, even within the same model (Abi-Rafeh et al., 2023). Furthermore, appropriate patient dispositions were suggested in only 56% of cases, and ChatGPT overestimated the urgency of patient needs in 44% of scenarios (Abi-Rafeh et al., 2023). These tendencies could lead to unnecessary healthcare utilization and heightened patient anxiety (Abi-Rafeh et al., 2023). While the overall performance of LLMs is encouraging, there is a clear consensus that these tools should not replace human expertise (Aliyeva et al., 2024; Boyd et al., 2024; Capelleras et al., 2024). This caution is particularly warranted given that most LLMs rely on pre-existing datasets and lack access to real-time, procedure-specific resources.

Given these limitations, this study sought to investigate whether Retrieval-Augmented Generation (RAG) could improve response accuracy by enabling models to access specific, credible rhinoplasty resources. Our findings indicated that while RAG has the potential to enhance the quality of AI responses, it also revealed significant challenges, including the retrieval of incomplete information, limited query interpretation, and difficulties in ensuring semantic consistency across the retrieved content. When responses were generated, they generally aligned well with the source text, suggesting that RAG could indeed enhance response accuracy when answers are produced. However, the primary limitation stemmed from a high rate of nonresponses, which impacted the mean accuracy scores. Including these nonresponses in our analysis was essential, as an unanswered question can be as unhelpful or misleading as an incorrect answer. This highlights that RAG, in its current form, may not yet be suitable as a standalone tool for postoperative care.

Our team identified that the answers to all prompts could be found within the source texts, therefore, the issue of nonresponse likely arose from challenges with context interpretation and phrase recognition. While the models handled straightforward questions effectively (e.g., “How should I sleep?”), they struggled with more nuanced queries such as “When can I get my face wet?” Although the source text referenced activities like showering and washing the face, the models failed to link these actions to the concept of “getting the face wet,” resulting in failure to retrieve documents from the vector database. This suggests that the current embedding model may have limitations in accurately capturing the nuanced semantics of these complex queries. Additionally, differences in phrasing between patient questions and source text likely also contributed to these performance limitations. To resolve this, prompts can be enhanced by offering clear instructions and extra context, helping to direct the model's understanding and improve the quality of its responses. In certain instances, relevant information was not retrieved due to incomplete indexing or dimensional constraints within the existing model. These factors emphasize the need for enhancements in the retrieval process to ensure more precise and contextually aware results.

Additional unique challenges with retrieval-augmented generation (RAG) have been identified in recent studies. Liu et al. found that model performance significantly decreases when relevant information is located in the middle of the text, as opposed to the beginning or end, potentially impacting retrieval accuracy (Liu et al., 2023). Improvements in chunking strategies, addressing chunk size, retrieval, and contextual reassembly, specifically with adaptive chunking, and advanced chunking mechanisms, may improve model performance. Furthermore, issues such as negative rejection (where models fail to produce an answer), difficulties with integrating information, and vulnerability to false information remain challenges for RAG-enhanced LLMs, although these models do exhibit some robustness to noise (Chen et al., 2023). Despite these limitations, RAG has shown promise as a valuable tool across various medical applications, including preoperative medicine (Ke et al., 2024), radiology report generation (Ranjit et al., 2023), and medical question answering (Xiong et al., 2024).

Building on these broader observations, our study identified distinct strengths and weaknesses among the models, offering valuable insights into their potential for assisting with patient education and support. Notably, Gemini-1.0-Pro-002 emerged as the most comprehensive model. This performance may stem from its design, which is particularly suited for handling complex natural language tasks and multiturn interactions, prioritizing detailed and thorough responses (Google Firebase, 2024, November 24). This emphasis on comprehensiveness likely comes at the expense of speed and adaptability. In contrast, Gemini-1.5-Flash-001, optimized for rapid responses,

demonstrated a focus on efficiency that may reduce the depth of its outputs (Google DeepMind, n.d.). PaLM 2, while designed for nuanced contextual understanding, logical reasoning, and conversational abilities, provided less actionable responses. This outcome is likely due to its prioritization of sophisticated language processing and conversational fluency over direct guidance (Google AI, n.d.-b). These findings illustrate the variability in model performance and underscore the importance of selecting tools tailored to the specific needs of clinical applications, such as balancing depth, clarity, and actionability.

Overall, these findings underscore both the promise and limitations of retrieval-augmented generation (RAG) in supporting postoperative patient care. While RAG models demonstrate potential as adjunctive tools by providing targeted, accurate information when responses are generated, the issues with nonresponses and context interpretation highlight areas needing further refinement. Moreover, clinical decision-making is inherently variable due to the subjective nature of human judgment, and RAG offers a promising avenue for creating greater consistency in decision-making processes, which could improve patient outcomes and reduce the risks associated with variability in human expertise. Although the literature highlights its potential, it is essential to address and systematically report RAG's limitations to ensure safe and effective integration into healthcare settings, including plastic and reconstructive surgery. Addressing these challenges is crucial for maximizing the clinical utility of RAG while minimizing the risks associated with its application.

#### *4.2 Future Directions*

To address the issues with Retrieval-Augmented Generation (RAG) found in this study, we plan to optimize the embedding model with domain-specific data and explore alternative similarity metrics to enhance semantic comprehension. We will also enhance our retrieval system by implementing context-aware techniques that consider the broader context of both queries and documents, ensuring that the retrieved results align with the query's intent beyond keyword relevance. Additionally, we intend to develop a hybrid retrieval approach that integrates traditional keyword-based search with vector-based embeddings, leveraging the strengths of both methodologies to maximize accuracy and relevance. These improvements aim to refine the system's ability to interpret complex queries effectively and deliver more precise, contextually aligned document retrieval.

Additional future work will focus on integrating advanced RAG techniques, such as proposition-based chunking to enhance semantic coherence, self-query retrieval to dynamically refine queries, and re-ranking methods to prioritize the most relevant and high-quality results. Furthermore, response readability and understandability must be improved to ensure responses are accessible and comprehensible to patients. Tailoring model training to recognize common patient language, interpret nuanced phrasing, and better link related concepts could address many of the limitations observed in this study.

Expanding RAG's application beyond rhinoplasty to encompass other high-demand procedures within plastic surgery, such as breast reconstruction, liposuction, and abdominoplasty, would also allow for a broader evaluation of its effectiveness across varied postoperative care scenarios. By refining these capabilities and applying RAG to a wider range of surgical contexts, future research can further establish the utility of RAG as a reliable, adaptable tool in plastic surgery patient management.

#### *4.3 Advantages and Limitations of this Study*

To our knowledge, this is the first study to assess the application of retrieval-augmented generation (RAG) for answering postoperative inquiries in the context of rhinoplasty care. This unique focus contributes to the growing body of research on AI in patient support by examining RAG's potential to provide accurate, tailored responses within a specific surgical setting. However, this study has several limitations that should be considered when interpreting the results.

First, we tested only four AI models, and it is possible that other models may perform differently in this setting. As AI technology rapidly evolves, models and algorithms will continue to improve, requiring future studies to stay current with advancements to ensure clinical relevance. Additionally, each question posed to the model in this study was prefaced with specific details about the type of rhinoplasty (open vs. closed, aesthetic vs. functional) and the timing of the procedure. In a real-world setting, patients may not provide such precise information when asking their postoperative questions, which could impact model performance and response accuracy. These simulated scenarios may limit the generalizability of the study.

Our study was also limited by language, as all questions were asked exclusively in English. This restriction does not account for the diverse linguistic backgrounds of patients who may require postoperative support in other languages, potentially affecting the generalizability of RAG's performance across multilingual patient populations.

Finally, our evaluation focused exclusively on questions related to open, aesthetic rhinoplasty, limiting the generalizability of our findings across other types of rhinoplasty or additional plastic surgery procedures. Future studies should aim to incorporate a broader range of procedural contexts and languages to comprehensively assess RAG's utility in diverse surgical scenarios.

## 5. Conclusions

While Retrieval-Augmented Generation (RAG) demonstrates the ability to provide accurate responses aligned with credible rhinoplasty resources, our findings indicate that it is not yet suitable as a standalone replacement for human expertise in postoperative care. Among the models evaluated, Gemini-1.0-Pro-002 proved to be the most comprehensive, whereas PaLM 2 scored lowest in actionability, highlighting variability in model performance. All models struggled with readability and understandability, pointing to a general need for improvements to ensure responses are accessible and comprehensible for patients.

The primary limitation observed in this study was the high rate of nonresponses, likely stemming from challenges in context interpretation and phrase recognition. These issues prevent RAG from consistently delivering the comprehensive and reliable support necessary for safe patient management. To move toward effective implementation in plastic surgery and other clinical settings, further refinement is needed to enhance RAG's ability to interpret nuanced questions, recognize diverse patient phrasing, and improve overall readability and understandability. Additionally, expanding research to evaluate RAG across various procedures and languages could provide a more complete understanding of its utility and limitations.

With targeted advancements, RAG has the potential to evolve into a valuable adjunctive tool in postoperative care, supporting healthcare professionals by addressing common patient inquiries and enhancing patient engagement while maintaining high standards of patient safety.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Supplementary File 1: Questions provided to AI Models; File 2: Model Responses using Retrieval-Augmented Generation.

**Author Contributions:** Conceptualization, Ariana Genovese, Srinivasagam Prabha and Antonio Jorge Forte; Formal analysis, Ariana Genovese, Srinivasagam Prabha, Sahar Borna and Cui Tao; Investigation, Ariana Genovese, Sahar Borna, Cesar Gomez-Cabello, Syed Ali Haider and Maissa Trabilisy; Methodology, Ariana Genovese and Srinivasagam Prabha; Project administration, Cui Tao and Antonio Jorge Forte; Software, Ariana Genovese and Srinivasagam Prabha; Supervision, Cui Tao and Antonio Jorge Forte; Validation, Ariana Genovese, Srinivasagam Prabha and Sahar Borna; Writing – original draft, Ariana Genovese and Srinivasagam Prabha; Writing – review & editing, Sahar Borna, Cesar Gomez-Cabello, Syed Ali Haider, Maissa Trabilisy, Cui Tao and Antonio Jorge Forte.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors upon request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Abi-Rafeh, J., Hanna, S., Bassiri-Tehrani, B., Kazan, R., & Nahai, F. (2023). Complications Following Facelift and Neck Lift: Implementation and Assessment of Large Language Model and Artificial Intelligence (ChatGPT) Performance Across 16 Simulated Patient Presentations. *Aesthetic Plastic Surgery*, 47(6), 2407-2414. <https://doi.org/10.1007/s00266-023-03538-1>
- Aliyeva, A., Sari, E., Alaskarov, E., & Nasirov, R. (2024). Enhancing Postoperative Cochlear Implant Care With ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support. *Cureus*, 16(2), e53897. <https://doi.org/10.7759/cureus.53897>
- American Society of Plastic Surgeons. (2023). 2023 Plastic Surgery Statistics Report. <https://www.plasticsurgery.org/documents/news/statistics/2023/plastic-surgery-statistics-report-2023.pdf>
- Borna, S., Gomez-Cabello, C. A., Pressman, S. M., Haider, S. A., & Forte, A. J. (2024). Comparative Analysis of Large Language Models in Emergency Plastic Surgery Decision-Making: The Role of Physical Exam Data. *Journal of Personalized Medicine*, 14(6), 612. <https://doi.org/10.3390/jpm14060612>
- Boroumand, M. A., Sedghi, S., Adibi, P., Panahi, S., & Rahimi, A. (2022). Patients' perspectives on the quality of online patient education materials: A qualitative study. *Journal of Education and Health Promotion*, 11. [https://doi.org/10.4103/jehp.jehp\\_1127\\_21](https://doi.org/10.4103/jehp.jehp_1127_21)
- Boyd, C. J., Hemal, K., Sorenson, T. J., Patel, P. A., Bekisz, J. M., Choi, M., & Karp, N. S. (2024). Artificial Intelligence as a Triage Tool during the Perioperative Period: Pilot Study of Accuracy and Accessibility for Clinical Application. *Plastic and Reconstructive Surgery – Global Open*, 12(2), e5580. <https://doi.org/10.1097/GOX.0000000000005580>
- Capelleras, M., Soto-Galindo, G. A., Cruellas, M., & Apaydin, F. (2024). ChatGPT and Rhinoplasty Recovery: An Exploration of AI's Role in Postoperative Guidance. *Facial Plastic Surgery : FPS*, 40(5), 628-631. <https://doi.org/10.1055/a-2219-4901>
- Chen, J., Lin, H., Han, X., & Sun, L. (2023). Benchmarking Large Language Models in Retrieval-Augmented Generation. *ArXiv e-prints*. <https://doi.org/10.48550/arXiv.2309.01431>
- Constantian, M. B. (2024). Closed technique rhinoplasty. In J. P. Rubin & P. C. Neligan (Eds.), *Plastic Surgery, Volume 2: Aesthetic Surgery* (Fifth Edition ed., pp. 607-646).
- Dwyer, T., Hoit, G., Burns, D., Higgins, J., Chang, J., Whelan, D., Kiroplis, I., & Chahal, J. (2023). Use of an Artificial Intelligence Conversational Agent (Chatbot) for Hip Arthroscopy Patients Following Surgery. *Arthroscopy, Sports Medicine, and Rehabilitation*, 5(2), e495-e505. <https://doi.org/10.1016/j.asmr.2023.01.020>
- Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Richard Andrew, T., & David, C. (2023). 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 Medical Education*, 9:e45312. <https://doi.org/10.2196/45312>
- Godin, M. S. (2012). Postoperative Care. In *Rhinoplasty: Cases and Techniques* (1st Edition ed., pp. 59-61). Thieme Medical Publishers. <https://doi.org/10.1055/b-002-85449>
- Gomez-Cabello, C. A., Borna, S., Pressman, S. M., Haider, S. A., Sehgal, A., Leibovich, B. C., & Forte, A. J. (2024). Artificial Intelligence in Postoperative Care: Assessing Large Language Models for Patient



- Recommendations in Plastic Surgery. *Healthcare (Basel, Switzerland)*, 12(11), 1083. <https://doi.org/10.3390/healthcare12111083>
- Google AI. (2024, November 21). *Gemini models*. <https://ai.google.dev/gemini-api/docs/models/gemini>
- Google AI. (n.d.-a). *The Gemini ecosystem* <https://ai.google/gemini-ecosystem>
- Google AI. (n.d.-b). *PaLM 2* <https://ai.google/discover/palm2>
- Google DeepMind. (n.d.). *Gemini Flash*. <https://deepmind.google/technologies/gemini/flash>
- Google Firebase. (2024, November 24). *Learn about the Gemini models* <https://firebase.google.com/docs/vertex-ai/gemini-models>
- Hoppe, I. C., Ahuja, N. K., Ingargiola, M. J., & Granick, M. S. (2013). A survey of patient comprehension of readily accessible online educational material regarding plastic surgery procedures. *Aesthetic Surgery Journal*, 33(3), 436-442. <https://doi.org/10.1177/1090820X13480013>
- Kahn, D. M., Rochlin, D. H., & Gruber, R. P. (2024). Secondary rhinoplasty. In J. P. Rubin & P. C. Neligan (Eds.), *Plastic Surgery, Volume 2: Aesthetic Surgery* (Fifth Edition ed., pp. 662-680).
- Kaschke, O. (2017). Postoperative Care and Management. In H. Behrbohm & E. Tardy (Eds.), *Essentials of Septorhinoplasty: Philosophy—Approaches—Techniques* (2nd Edition ed., pp. 248-256). Georg Thieme Verlag KG. <https://doi.org/10.1055/b-005-143328>
- Ke, Y., Jin, L., Elangovan, K., Abdullah, H. R., Liu, N., Sia, A. T. H., Soh, C. R., Tung, J. Y. M., Ong, J. C. L., & Ting, D. S. W. (2024). Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report. *ArXiv e-prints*. <https://doi.org/10.48550/arXiv.2402.01733>
- Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, 388(13), 1233-1239. <https://doi.org/10.1056/NEJMSr2214184>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv e-prints*. <https://doi.org/10.48550/arXiv.2005.11401>
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the Middle: How Language Models Use Long Contexts. *ArXiv e-prints*. <https://doi.org/10.48550/arXiv.2307.03172>
- Mantelakis, A., Assael, Y., Sorooshian, P., & Khajuria, A. (2021). Machine Learning Demonstrates High Accuracy for Disease Diagnosis and Prognosis in Plastic Surgery. *Plastic and Reconstructive Surgery – Global Open*, 9(6), e3638. <https://doi.org/10.1097/GOX.00000000000003638>
- Nguyen, T. P., Carvalho, B., Sukhdeo, H., Joudi, K., Guo, N., Chen, M., Wolpaw, J. T., Kiefer, J. J., Byrne, M., Jamroz, T., Mootz, A. A., Reale, S. C., Zou, J., & Sultan, P. (2024). Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia. *BJA Open*, 10. <https://doi.org/10.1016/j.bjao.2024.100280>
- The Patient Education Materials Assessment Tool (PEMAT) and User's Guide. (2024). In.
- Ranjit, M., Ganapathy, G., Manuel, R., & Ganu, T. (2023). Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models. *ArXiv e-prints*. <https://doi.org/10.48550/arXiv.2305.03660>
- Rohrich, R. J., & Afrooz, P. N. (2024). Open technique rhinoplasty. In J. P. Rubin & P. C. Neligan (Eds.), *Plastic Surgery, Volume 2: Aesthetic Surgery* (Fifth Edition ed., pp. 580-606). Elsevier.
- Shoemaker, S. J., Wolf, M. S., & Brach, C. (2014). Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Education and Counseling*, 96(3), 395-403. <https://doi.org/10.1016/j.pec.2014.05.027>
- Small, W. R., Wiesenfeld, B., Brandfield-Harvey, B., Jonassen, Z., Mandal, S., Stevens, E. R., Major, V. J., Lostraglio, E., Szerencsy, A., Jones, S., Aphinyanaphongs, Y., Johnson, S. B., Nov, O., & Mann, D. (2024).

- Large Language Model-Based Responses to Patients' In-Basket Messages. *JAMA Network Open*, 7(7), e2422399. <https://doi.org/10.1001/jamanetworkopen.2024.22399>
- Stiller, C., Brandt, L., Adams, M., & Gura, N. (2024). Improving the Readability of Patient Education Materials in Physical Therapy. *Cureus*, 16(2). <https://doi.org/10.7759/cureus.54525>
- Tabbal, G. N., & Matarasso, A. (2021). Rhinoplasty. In B. A. Mast (Ed.), *Plastic Surgery: A Practical Guide to Operative Care* (1st Edition ed., pp. 203-214). Thieme <https://doi.org/10.1055/b-006-163728>
- Xiong, G., Jin, Q., Lu, Z., & Zhang, A. (2024). Benchmarking Retrieval-Augmented Generation for Medicine. *ArXiv e-prints*. <https://doi.org/10.48550/arXiv.2402.13178>

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.