


DOI: <https://doi.org/10.5554/22562087.e1092>

ChatGPT's learning and reasoning capacity in anesthesiology

Capacidad de aprendizaje y razonamiento de ChatGPT en temas de anestesiología

Gustavo Cruz^a , Santiago Pedroza^b , Fredy Ariza^a ^aAnesthesiology Department, Fundación Valle del Lili. Cali, Colombia.^bClinical Research Center, Fundación Valle del Lili. Cali, Colombia.**Correspondence:** Departamento de Anestesiología, Fundación Valle del Lili, Cra. 98 # 18-49. Cali, Colombia.**E-mail:** gustavo.cruz@fvl.org.co

What do we know about this problem?

ChatGPT's medical reasoning has been tested in parasitology tests and in the United States Medical Licensing Examination questions. Moreover, ChatGPT has been shown to be a viable tool for radiological decision-making, potentially improving clinical workflows. However, the use of ChatGPT in clinical decision-making in real cases is limited and its application in the anesthesiology field is still unknown.

How does this study contribute?

The study showed acceptable accuracy in the basic knowledge test, high relevance for the management of specific difficult airway clinical cases, and the ability to improve after training. This study highlights the potential of large language artificial intelligence models (LLMs) and their possible applications in anesthesiology.

How to cite this article

Cruz G, Pedroza S, Ariza F. ChatGPT's learning and reasoning capacity in anesthesiology. Colombian Journal of Anesthesiology. 2024;52:e1092.

Abstract

Introduction: Over the past few months, ChatGPT has raised a lot of interest given its ability to perform complex tasks through natural language and conversation. However, its use in clinical decision-making is limited and its application in the field of anesthesiology is unknown.

Objective: To assess ChatGPT's basic and clinical reasoning and its learning ability in a performance test on general and specific anesthesia topics.

Methods: A three-phase assessment was conducted. Basic knowledge of anesthesia was assessed in the first phase, followed by a review of difficult airway management and, finally, measurement of decision-making ability in ten clinical cases. The second and the third phases were conducted before and after feeding ChatGPT with the 2022 guidelines of the American Society of Anesthesiologists on difficult airway management.

Results: On average, ChatGPT succeeded 65% of the time in the first phase and 48% of the time in the second phase. Agreement in clinical cases was 20%, with 90% relevance and 10% error rate. After learning, ChatGPT improved in the second phase, and was correct 59% of the time, with agreement in clinical cases also increasing to 40%.

Conclusions: ChatGPT showed acceptable accuracy in the basic knowledge test, high relevance in the management of specific difficult airway clinical cases, and the ability to improve after learning.

Keywords: ChatGPT; Artificial intelligence; Anesthesiology; Difficult airway; Learning; Reasoning; Decision-making.

Resumen

Introducción: En los últimos meses, ChatGPT ha suscitado un gran interés debido a su capacidad para realizar tareas complejas a través del lenguaje natural y la conversación. Sin embargo, su uso en la toma de decisiones clínicas es limitado y su aplicación en el campo de anestesiología es desconocido.

Objetivo: Evaluar el razonamiento básico, clínico y la capacidad de aprendizaje de ChatGPT en una prueba de rendimiento sobre temas generales y específicos de anestesiología.

Métodos: Se llevó a cabo una evaluación dividida en tres fases. Se valoraron conocimientos básicos de anestesiología en la primera fase, seguida de una revisión del manejo de vía aérea difícil y, finalmente, se midió la toma de decisiones en diez casos clínicos. La segunda y tercera fases se realizaron antes y después de alimentar a ChatGPT con las guías de la Sociedad Americana de Anestesiólogos del manejo de la vía aérea difícil del 2022.

Resultados: ChatGPT obtuvo una tasa de acierto promedio del 65 % en la primera fase y del 48 % en la segunda fase. En los casos clínicos, obtuvo una concordancia del 20 %, una relevancia del 90 % y una tasa de error del 10 %. Posterior al aprendizaje, ChatGPT mejoró su tasa de acierto al 59 % en la segunda fase y aumentó la concordancia al 40 % en los casos clínicos.

Conclusiones: ChatGPT demostró una precisión aceptable en la prueba de conocimientos básicos, una alta relevancia en el manejo de los casos clínicos específicos de vía aérea difícil y la capacidad de mejorar secundaria a un aprendizaje.

Palabras clave: ChatGPT; Inteligencia artificial; Anestesiología; Vía aérea difícil; Aprendizaje; Razonamiento; Toma de decisiones.

INTRODUCTION

Interest in the development and application of artificial intelligence (AI) has been growing in recent years. This new technology has transformed the way in which we approach various tasks, including data analysis, industrial automation and virtual assistance, among others (1). The exponential evolution in data storage capacity and the growth of information digitalization set the stage for AI's main function, namely, enabling large dataset analyses and pattern identification, as well as providing answers that would normally require the involvement of human intelligence to be completed (2).

AI has had a significant impact on many sectors, including development, finance, as well as humanistic and scientific work. Although its development in the field of medical sciences is promising, its application in clinical care continues to be limited (3). In the clinical realm, its use is gradually growing, especially in text generation. However, the increase in non-structured text fields, together with the lack of interoperability and synergistic communications between AI technology systems and health infrastructure results

in a substantial shortage of data with adequate structure and legibility that can be assimilated by the AI systems required to conceive and develop deep learning algorithms (3).

Moreover, decision-making regarding clinical actions — usually dependent on multiple factors — also makes the application of this technology difficult. AI has been used in some medical specialties, albeit with no far-reaching achievements. A review of 23 studies on AI application for breast cancer detection conducted between 2010 and 2018 showed that most of the studies were retrospective and small in size, with no possibility of generalizing the results (4). Furthermore, a systematic review conducted in 2021 concluded that current evidence is insufficient to support AI implementation in early breast cancer detection (5).

ChatGPT (Chat Generative Pre-Trained Transformer) is a natural language processing model based on the architecture of the GPT-3.5 language model which has created great interest because of its ability to generate, understand and interpret human language by means of information technology systems (6). Developed by Open AI, the artificial intelligence research company out of San Francisco, California, it

was trained by a variant of the Transformer architecture, a neural network deep learning model designed to handle sequential data. A dataset of 40 GB of text was used, resulting in a model with 1.5 billion parameters (7,8). ChatGPT is the latest GPT-3 variant, and it is specifically designed to interact with the user (7). However, risks associated with its implementation in healthcare have been identified, including bias, data confidentiality, misleading information and lack of adequate referencing (9).

Huh assessed ChatGPT's performance in a parasitology test and compared it with the test performed by medical students in Korea. ChatGPT was correct 60.8% of the time, compared to the 90.8% mean of 77 students (10). Kung et al., assessed ChatGPT's performance in questions of the United States Medical Licensing Examination (USMLE) and found evidence of understandable reasoning ability and valid clinical ideas (3). ChatGPT obtained a score of more than 60% in the NBME-Free-Step-1 dataset, equivalent to a passing grade for a third-year medical student (3). On the other hand, Rao et al. demonstrated ChatGPT's viability as a tool for radiological decision-making, potentially enhancing clinical workflows (11).

The use of ChatGPT in clinical decision-making in actual cases is limited and its application in the anesthesiology field is unknown. The objective of this study was to assess ChatGPT’s basic and clinical reasoning, as well as its ability to learn in a performance test on general and specific anesthesiology topics.

METHODS

ChatGPT: A natural language processing system that uses a transformer-type neural network to generate coherent and relevant responses. The model learns patterns in large text datasets, allowing it to capture contextual and syntactic information from the input text and thus generate accurate responses (12). Likewise, ChatGPT can adjust and adapt to different domains and tasks through specific data-based training. However, the current ChatGPT model is not capable of searching the Internet when generating answers but rather resorts to patterns in its training data (11).

Input source: Four information sources were selected: 1) A bank of review questions for the primary Fellowship of the Royal College of Anaesthetists (FRCA) examination. The primary FRCA is a

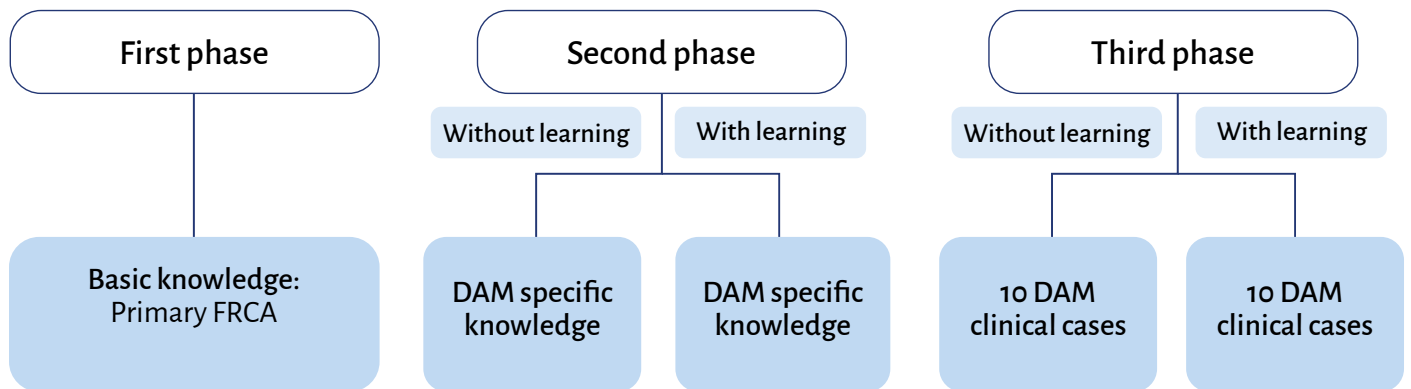
graduate examination that anesthetists in training in the United Kingdom must pass before applying to Higher Specialized Anesthesia Training (13). The examination consists of 45 true or false multiple choice questions (MTF) and 45 single best answer (SBA) multiple choice questions. The rate of right answers required to pass the exam changes every year but normally ranges between 58% and 70% (14). A total of 840 questions were pulled from three books of MTF question banks used to prepare for the test (Qbase anesthesia: 1, 2 and 3) (15-17); 2) Specialized difficult airway management (DAM) exam. The questions were taken from Anesthesia HUB (18), a central web-based anesthesiology source in which only the DAM questions were filtered, retrieving 44 SBA questions; 3) Ten actual clinical case reports on DAM taken from PubMed and selected by the researchers (19-23); 4) 2022 practice guidelines of the American Society of Anesthesiologists (ASA) for difficult airway management (24).

Assessment method: The assessment was carried out in three phases (Figure 1). In the first phase, the primary FRCA knowledge was assessed using 11 tests containing 840 MTF questions in which the statement presented 5 choices. The second phase consisted of a specialized DAM exam using SBA questions. Finally,

the third phase assessed 10 real difficult airway management clinical cases in which questions asked about the best option to secure the airway and the anesthetics to be administered. The second and the third phase were carried out before and after feeding ChatGPT with the ASA 2022 DAM guidelines.

Model coding/input: Coding was organized in four sections: 1) True or false prompts. The text entry of the question was made using the statement “The following question has 5 answer choices. Answer true or false for each choice;” 2) Multiple choice questions. Single answer prompts. The text input of the question was made with the heading “Choose only one answer (the best choice) for the following question;” 3) Clinical cases. An “open-ended request” or “unrestricted request” was used for this type of assessment. In this coding, all answer choices were removed and an interrogative phrase was used in the heading. Two input questions were asked after presenting the clinical case: “In your opinion, after reading clinical case 1, What is the best option (choose only one) to secure the airway? 2. Which would be the best choice of anesthetics to secure the airway? 4) Feeding difficult airway management guidelines. The 2022 ASA guidelines for DAM in written form (the

Figure 1. ChatGPT assessment phases.



FRCA: Bank of review questions for the primary Fellowship of the Royal College of Anaesthetists examination.

DAM: Difficult airway management.

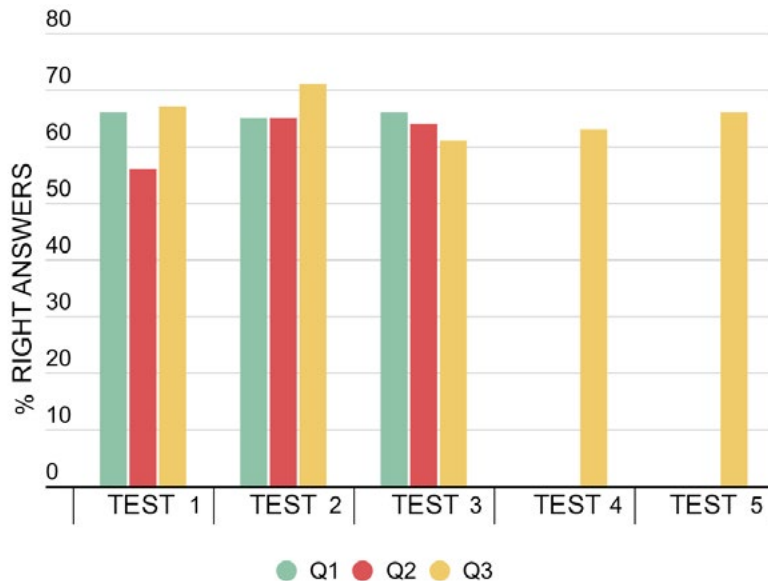
Source: Authors.

Table 1. ChatGPT's performance in basic anesthesia knowledge assessment.

Q1 Questions bank				
Test	Right	Wrong	Total	Success rate (%)
Test 1 (90)	295	155	450	66
Test 2 (90)	291	159	450	65
Test 3 (90)	295	155	450	66
Q2 Questions bank				
Test	Right	Wrong	Total	Success rate (%)
Test 1 (90)	256	194	450	56.9
Test 2 (90)	296	154	450	65.8
Test 3 (90)	292	158	450	64.9
Q3 Questions bank				
Test	Right	Wrong	Total	Success rate (%)
Test 1 (60)	201	99	300	67.0
Test 2 (60)	215	85	300	71.7
Test 3 (60)	184	116	300	61.3
Test 4 (60)	191	109	300	63.7
Test 5 (60)	200	100	300	66.7

Source: Authors.

Figure 2. ChatGPT's performance in basic anesthesia knowledge assessment.



Source: Authors.

parts that contained images or algorithms were transcribed to text) were adapted, and then the guidelines were incorporated step by step (given ChatGPT's transcription word limit per message), starting always with the statement: "Learn this information and take it into account for future questions."

Clinical case assessments: A rating method dividing the answers into three groups was created: concordant, relevant and incorrect. Concordant answers were the same as the first-line or the most adequate management selected also in the actual clinical case. Relevant answers were those

that corresponded to any acceptable management. Finally, incorrect responses were those related to the wrong treatment options.

Given that a concordant answer is also coherent (because first-line management was also part of the acceptable management category), all concordant answers were also classified as coherent.

Answers to clinical cases were assessed independently by two anesthesiologists who were unaware of the purpose of the research or each other's responses. Care was taken to ensure that the two physicians had no knowledge of their respective assessments, and agreement between the two examiners in terms of category assignment was examined.

RESULTS

To assess basic anesthesia knowledge, 840 questions divided into 11 tests were asked. Each question contained 5 true or false statements, for a total of 4200 answers. In this initial assessment, ChatGPT chose the right answer 65% of the time (Table 1, Figure 2). In the next exam, using specific SBA questions pertaining to difficult airway management, ChatGPT had a 48% success rate (Table 2).

In the last assessment, the answers to the two questions on each clinical case were evaluated and categorized by two anesthesiologists through inspections of the explanatory content, with 100% agreement. ChatGPT provided answers and explanations with 20% agreement, 90% relevance and 10% error rate.

Once the agreement and relevance of the answers were determined, ChatGPT's ability to learn was assessed based on the change in its answers after learning with the ASA 2022 guidelines for DAM (18). In the MTF test, ChatGPT improved its success rate to 59% (Table 2). Likewise, in terms of the answers to the clinical cases, ChatGPT increased agreement to 40% (Figure 3).

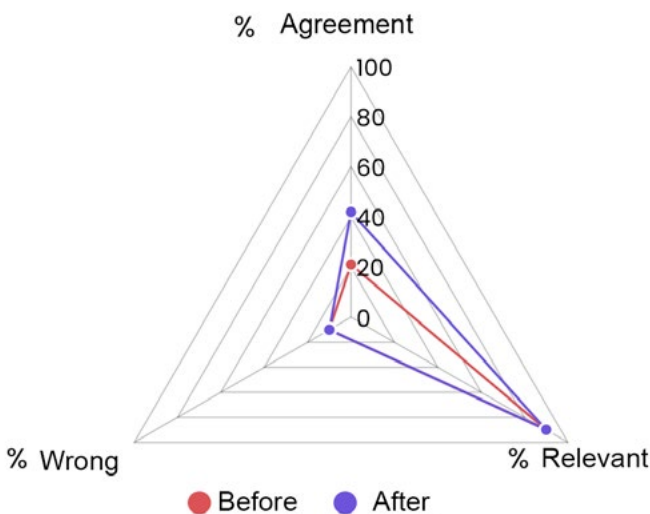
Table 2. ChatGPT's performance in specific anesthesia knowledge assessment.

Specific airway management test				
Learning	Right	Wrong	Total	Per cent success
BL	21	23	44	48
AL	26	18	44	59

BL: before learning, **AL:** after learning.

Source: Authors.

Figure 3. ChatGPT agreement, relevance and fail rates in the questions pertaining to clinical cases before and after learning.



Source: Authors.

DISCUSSION

This study showed evidence of the potential and ability of language models to make decisions at different levels of complexity in anesthesia. ChatGPT showed acceptable accuracy in the basic knowledge test, high relevance in specific difficult airway management clinical cases, and the ability to improve after learning.

Prior studies have examined ChatGPT's reasoning accuracy when it comes to medical questions. Older models (GPT3) were shown to have a 36.7% performance in the United States Medical Licensing Examination(25). Few months later,

following the update of the GPT3 version, Kung et al. demonstrated accuracy above 50% in all USMLE exams, coming close to the passing grade (approximately 60%), and showing acceptable basic and clinical reasoning (3). This study reports similar accuracy in the MTF for the basic anesthesiology exam, with a score (65%) within the upper range of concordant answer rates required to pass the primary FRCA examination (58-70%).

Regarding the specific SBA test on DAM, ChatGPT's performance did not reach the passing threshold (higher than 60%). One possible explanation for this result may be the way in which the questions were

designed. Given that difficult airway is an anesthetic condition for which there are multiple treatment options, included in the different answer options, asking for a single best answer may have lowered the success rate. This situation occurred again in the last test in which, despite a 20% agreement rate in open-ended questions on clinical cases, the rate of relevant answers, i.e., a valid answer other than the first-line choice, was 90%.

Similar results were obtained by Yeo et al. when they assessed ChatGPT's performance with questions on cirrhosis and hepatocellular carcinoma. They reported 75% complete and comprehensible answers in the basic knowledge, treatment and lifestyle categories. However, this percentage dropped to 50% in the diagnosis category. Both the Yeo et al. as well as this research point to a sound knowledge base in ChatGPT's responses, albeit with a limited ability in all cases to provide the best individualized recommendations (valid but not the best possible answers). Therefore, the recommendation at this point is to use it only as a complementary information tool (26).

ChatGPT's ability to learn was demonstrated, with improved success rate in the SBA-type test and also in terms of agreement in the clinical cases after feeding it with the DMA guidelines. This finding highlights two important points: first, ChatGPT's and other models' huge potential for improvement once barriers to their effective feeding of information diminish; and second, the indication that ChatGPT's inaccuracies can be attributed more to information deficiencies than to processing errors. This premise is supported by the high relevance observed both before and after the learning process, and also by improved agreement after learning. The latter was also described by Kung et al., who found that inaccurate answers were driven mainly by missing information, resulting in reduced understanding and AI indecision (3).

This study has important limitations. Input data were relatively scant, particularly for the assessment of real clinical cases.

This could have affected the depth and scope of the analyses. Moreover, a stronger study of AI failure mode (e.g., asking for a more detailed rationale for each answer and assessing analysis errors) could provide valuable information on the etiology of the inaccuracies and disagreements. Also, some limitations stem from restrictions inherent to ChatGPT. These include the inability to search for new/recent information on the Internet and to attribute factual information to one source. These limitations must be taken into account when assessing this tool's clinical decision-making.

It is evident that, more and more, AI-based technologies will become part of daily life, including their use in future tools specializing in clinical decision-making. This work is of cardinal importance in this process, as an entry door to potential uses in anesthesiology. The weaknesses of the study and of ChatGPT, subject to improvement in future releases of this new technology, are underscored.

CONCLUSIONS

ChatGPT offers acceptable accuracy in basic knowledge tests, high relevance in the management of specific difficult airway clinical cases, and the ability to improve after learning. This study highlights the potential and possible uses of language models and of AI in anesthesiology. Strategies to minimize risks and obtain the best benefits should be developed.

ETHICAL RESPONSIBILITIES

Ethics committee endorsement

The study did not involve animals or humans, hence the absence of a statement on ethical approval in the Methods section.

Protection of human and animal subjects

The authors declare that no experiments were performed on animals for this study.

The authors declare that the procedures followed were in accordance with the ethical standards of the Responsible Human Experimentation Committee and of the World Medical Association and the Declaration of Helsinki.

Data confidentiality

The authors declare that they have followed the protocols of their work center on the publication of patient data.

Right to privacy and informed consent

The authors declare that no patient data are disclosed in this article. The authors obtained informed consent from the patients and/or subjects referred to in this article. This document is available from the corresponding author.

ACKNOWLEDGEMENTS

Authors' contributions

GC: Conceptualization, research, manuscript writing, review and editing, supervision.

SP: Conceptualization, research, manuscript writing, review and editing, tables 1-2 and figures 1-3.

FA: Research, manuscript writing, review and editing.

Assistance for the study

None declared.

Funding and sponsorship

None declared.

Conflict of interest

None declared.

Presentations

None declared.

Appreciation

None declared.

REFERENCES

- Kim SW, Kong JH, Lee SW, Lee S. Recent advances of artificial intelligence in manufacturing industrial sectors: A review. *Int J Precision Engin Manufactur.* 2022;23:111-29. doi: <https://doi.org/10.1007/s12541-021-00600-3>.
- Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emergency Medicine Australasia EMA.* 2018;30:870-4. doi: <https://doi.org/10.1111/1742-6723.13145>.
- Kung TH, Cheatham M, Medenilla A, Sillos C, De León L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health.* 2023;2:e0000198. doi: <https://doi.org/10.1371/journal.pdig.0000198>.
- Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices.* 2019;16:351-62. doi: <https://doi.org/10.1080/17434440.2019.1610387>.
- De Vries CF, Colosimo SJ, Boyle M, Lip G, Anderson LA, Staff RT, et al. AI in breast screening mammography: breast screening readers' perspectives. *Insights Imaging.* 2022;13. doi: <https://doi.org/10.1186/s13244-022-01322-4>.
- Stokel-Walker C, Noorden R. What ChatGPT and generative AI mean for science. *Nature.* 2023;214-6. doi: <https://doi.org/10.1038/d41586-023-00340-6>.
- Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J Med Syst.* 2023;47. doi: <https://doi.org/10.1007/s10916-023-01925-4>.
- Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you

- need. 31st Conference on Neural Information Processing Systems, 2017.
9. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *healthcare (Basel)* 2023;11. doi: <https://doi.org/10.3390/healthcare11060887>.
 10. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof.* 2023;20:1. doi: <https://doi.org/10.3352/jeehp.2023.20.1>.
 11. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD, et al. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. doi: <https://doi.org/10.1101/2023.02.02.23285399>.
 12. Natalie. Open AI - What is ChatGPT? [internet]. 2023 [citado: 2023 abr 26]. Disponible en: <https://help.openai.com/en/articles/6783457-what-is-chatgpt>
 13. Royal College of anaesthetists. FRCA Primary MCQ Examination. n.d.
 14. The Candidate The newsletter for FRCA candidates. n.d.
 15. Blunt M, Hammond E, McIndoe A. Qbase anaesthesia. Vol. 1, MCQs for the anaesthesia primary. Greenwich Medical Media; 1997.
 16. Blunt M, Hammond E, McIndoe A. Qbase anaesthesia. Vol. 2, MCQs for the final FRCA. Greenwich Medical Media; 1997.
 17. Hammond E, McIndoe A. Qbase anaesthesia. Vol. 3, MCQs in medicine for the FRCA. Greenwich Medical Media; 1999.
 18. Moss D. Anesthesia HUB. EXAMS [internet]. 2013 [citado: 2023 abr 26]. Disponible en: <https://www.anesthesiahub.com/>.
 19. Hariharasudhan B, Mane R, Gogate V, Dhorigol M. Successful management of difficult airway: A case series. *J Scient Soc.* 2016;43:151. doi: <https://doi.org/10.4103/0974-5009.190547>.
 20. Li M, Zhang L. Management of unexpected difficult airway in perioperative period: A case report. *Asian J Surg.* 2021;44:1564-5. doi: <https://doi.org/10.1016/j.asjsur.2021.08.041>.
 21. Pai Bh P, Shariat AN. Revisiting a case of difficult airway with a rigid laryngoscope. *BMJ Case Rep.* 2019;12. doi: <https://doi.org/10.1136/bcr-2018-224616>.
 22. Rugnath N, Rexrode LE, Kurnutala LN. Unanticipated difficult airway during elective surgery: A case report and review of literature. *Cureus.* 2022. doi: <https://doi.org/10.7759/cureus.32996>.
 23. González-Benito E, Del Castillo Fernández De Betoño T, Pardos PC, Ruiz PE. Difficult airway in a patient with lymphoma. A case report. *Rev Española Anestesiología Reanim.* 2021;68:297-300. doi: <https://doi.org/10.1016/j.redare.2020.05.024>.
 24. Apfelbaum JL, Hagberg CA, Connis RT, Abdelmalak BB, Agarkar M, Dutton RP, et al. American Society of Anesthesiologists Practice Guidelines for Management of the Difficult Airway. *Anesthesiology.* 2022;136:31-81. doi: <https://doi.org/10.1097/ALN.0000000000004002>.
 25. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams 2020. doi: <https://doi.org/10.3390/app11146421>.
 26. Hui Yeo Y, Samaan JS, Han Ng W, Ting P-S, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* 2023. doi: <https://doi.org/10.3350/cmh.2023.0089>.

Supplemental material. ChatGPT agreement, relevance and failure in the questions pertaining to difficult airway clinical cases.

Cases	Question	Before learning			After learning		
		Agreement	Relevante	Wrong	Agreement	Relevante	Wrong
1	1	1	1	0	1	1	0
	2	0	1	0	0	1	0
2	1	0	1	0	1	1	0
	2	0	1	0	1	1	0
3	1	0	1	0	1	1	0
	2	0	1	0	0	1	0
4	1	0	0	1	0	0	1
	2	0	0	1	0	0	1
5	1	1	1	0	0	1	0
	2	0	1	0	0	1	0
6	1	0	1	0	1	1	0
	2	0	1	0	0	1	0
7	1	0	1	0	0	1	0
	2	0	1	0	0	1	0
8	1	0	1	0	0	1	0
	2	1	1	0	1	1	0
9	1	0	1	0	0	1	0
	2	0	1	0	0	1	0
10	1	1	1	0	1	1	0
	2	0	1	0	1	1	0
Total	20	4	18	2	8	18	2

Source: Authors.